

Integration of human and machine translation

Marcello Federico

Fondazione Bruno Kessler

MT Marathon, Prague, Sept 2013

Motivation

- **Human translation (HT)**
 - worldwide demand for translation services has accelerated, due to globalization and growth of the Information Society
- **Gap between MT and HT**
 - MT has improved significantly but independently from HT
 - MT research has not directly addressed how to improve HT
 - Today professional translators barely use MT
- **The unavoidable adoption of MT**
 - Post-editing experiments have shown great promise
 - Integration of HT and MT is still an open problem!

Questions

- How do human translators work?
- What tools do they use?
- How is productivity measured?
- How can MT help human translators?
- What are important problems to solve?
- Why should MT researchers care?
- ...

Outline

- Typical translation-industry workflow
- Computer assisted translation tools
- Simple MT-CAT integration
- Two undergoing research projects
 - their research challenges
 - their experimental platforms

Scenario



All our translators
got a CAT tool!



Scenario



Computer Assisted Translation (CAT) is the dominant technology in the translation industry

CAT tools: special text editors supporting many document formats and integrating information from different sources.

CAT Tools

Source/target text is split into *segments*

Translation progresses segment by segment

Provides helps from different sources:

- spell checkers
- dictionaries
- terminology managers
- concordancers
- *translation memory (TM)*
- and recently *machine translation (MT)*

CAT Tool

SDL Translation Management System - VDEVOTETMS06 - / Inbox / Demo Environment / SDL Product Overview - Windows Internet Explorer

SDL Translation Management System

/ Inbox / Demo Environment / SDL Product Overview (Translation | EN-US > DE) - [SDL Product Overview.docx]

Back Refresh Help Info Print Log Out Save Revert Submit Reject QA Check Override

Preview Task Comments Return To Inbox TM

File Home Segment Format Tools Actions

Confirm Translation Cut Undo Copy Redo Paste Find and replace Concordance search Go to segment Previous segment Next segment

Display: All segments Containing: All segments

General

Segment Types

- New translated content
- 100% matches
- Fuzzy matches
- Unconfirmed
- Not translated
- Draft
- Duplicates

Segment Review

- With comments

Segment Locking

- Locked
- Unlocked

Segment Content

- Number only

When new content is written and submitted for translation SDL TMS automatically checks the content against previously translated content using the latest patented technology and advanced linguistic processing.

Segment ID	Source Text	Match %	Target Text
1	When new content is written and submitted for translation SDL TMS automatically checks the content against previously translated content using the latest patented technology and advanced linguistic processing.	100%	Wenn neue Inhalte automatisch gegen bereits übersetzte Inhalte mittels neuer Sprachverarbeitungstechnologie überprüft werden.
5900	Enables corporations to centralise all multilingual assets into a centralised repository.	100%	SDL TMS ermöglicht es Unternehmen, alle mehrsprachigen Inhalte in einer zentralisierten Datenbank zu verwalten.
5901	When new content is written and submitted for translation SDL TMS automatically checks the content against previously translated content using the latest patented technology and advanced linguistic processing.	100%	Wenn neue Inhalte automatisch gegen bereits übersetzte Inhalte mittels neuer Sprachverarbeitungstechnologie überprüft werden.
5902	Any content matched is delivered back translated, whilst new content requiring translation is automatically delivered down into the translation supply chain for human translation.	100%	Bereits übersetzte Inhalte werden automatisch zurück übersetzt, während neue Inhalte, die eine Übersetzung erfordern, automatisch in den Übersetzungsprozess gegeben werden.
5903	For more information about SDL TMS please visit our translation management section.	100%	Weitere Informationen über SDL TMS finden Sie in der Rubrik „Translation Management“.
5904	SDL Knowledge-based Translation System (SDL KbTS™)		SDL Knowledge-based Translation System (SDL KbTS™)
5905	Provides high-quality translations, accelerated time-to-market and reduced total cost for the world's leading brands.	82%	SDL KbTS™ liefert führenden Unternehmen weltweit qualitativ hochwertige Übersetzungen, beschleunigt die Time-to-Market und ermöglicht eine Reduzierung der Gesamtkosten.
5906	The power of the solution lies in the combination of sophisticated machine translation technology with other translation automation	100%	Der Vorteil der Lösung liegt in der Kombination hochentwickelter maschineller Übersetzungstechnologie mit weiteren automatisierten

Context: TEXT Source File: C:\Documents and Settings\Administrator\Desktop\SDL Product Overview.docx Source TM: DemoTM

/2010 10:44:38 AM dev_tms_srvc_usr

TU(s): 11239 INS

Local intranet | Protected Mode: Off 100%

Vanilla CAT Tool

The screenshot displays the Vanilla CAT Tool interface, which is divided into several functional areas:

- Menu Bar:** Contains 'File', 'Edit', 'View', and 'Help'.
- Editor:** The central workspace for text. It contains three paragraphs:
 - Paragraph 1: "But I must explain to you how all this mistaken idea of denouncing pleasure and praising pain was born and I will give you a complete account of the system, and expound the actual teachings of the great explorer of the truth, the master-builder of human happiness."
 - Paragraph 2: "Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt."
 - Paragraph 3: "No one rejects, dislikes, or avoids pleasure itself, because it is pleasure..."A yellow sticky note with a red tab is placed over the end of the third paragraph, containing the text: "Suggestion by MT or TM, ranked by a dynamic suggestion manager."
- TM Matches / MT Suggestion:** A panel on the right side, containing a yellow sticky note with a red tab that reads: "TM matches and partial MT fragments (informative MT)."
- Concordance / Terminology:** A panel on the right side, containing a yellow sticky note with a red tab that reads: "Terminology is automatically extracted from the MT phrase-table (self-tuning and informative MT)."
- Chat:** A panel on the right side, containing a yellow sticky note with a red tab that reads: "Collaboration between customers, translators and MT provider."
- Project Analytics:** A status bar at the bottom of the editor area, displaying: "Project Analytics: % completion; % repetitions; % fuzzy matches; % new words"

Terminology

- **Terms:** words and compound words that in specific contexts have specific meanings.
- **Termbase:** database consisting of terms and related information, usually in multilingual format.

Terminology

Terminology database

Term: Domain:

Source: Target:

Domain: LAW	Italiano
Term	concorrenza sleale
Reliability	3 (reliable)
Term reference	Enc Giuridica,Treccani,Roma,vol.VII,1988,s.v.concorrenza II;Codice Civile art.2598
Date	29/09/2009
	English
Definition	an attempt to do better than another company by using techniques which are not fair,such as importing foreign goods at very low prices or by wrongly criticising a competitor's products
Definition reference	3 (Dict of Accounting,Collin-Joliffe,1992)
Term	unfair competition
Date	29/09/2009

Concordance

- **Concordance:** occurrence of a word in a texts together with its context.
- Bilingual concordances show use of words in parallel texts.

Concordancer

Bilingual concordance

Source: EN-English Target: rabbit

Search string: EQUAL TO

Select corpus: Alice in Wonderland

She felt very sleepy, when suddenly a White **rabbit** with pink eyes ran close by her.

nor did Alice think it so unusual to hear the **rabbit** say to itself "Oh dear! Oh dear! I shall be too late!"

But when the **rabbit** actually took a watch out of its waistcoat-pocket, and looked at it, and then hurried on, Alice started to her feet, for she remembered that she had never before seen a rabbit with either a waistcoat-pocket or a watch to take out of it, and she ran across the field after it, and was just in time to see it pop down a large rabbit-hole under the hedge.

The **rabbit**-hole went straight on like a tunnel for some way, and then dipped suddenly down, so suddenly that Alice had no time to think about stopping herself before she found herself falling down what seemed to be a very deep well.

她感到昏昏欲睡，就在此时，一只长着粉红色眼睛的白兔突然打她身边跑过。

A爱丽丝听见那兔子自言自语地说：“哎呀！哎呀！我要迟到了！”她也不认为这有什么异常。

然而当兔子居然从背心口袋中掏出一只表，瞧了瞧，然后又匆匆赶路时，爱丽丝才一跃而起，因为在她的记忆中，从来没见过兔子穿有口袋的背心，也没见过兔子从口袋里掏出一只表来。于是她跟在那兔子后面跑过田野，刚好来得及看见它一下子钻进树篱下的一个大兔子洞。

兔子洞像隧道一样往前延伸，随后就忽地往下方拐去，这一拐来得太突然了，爱丽丝还来不及想停住，就发现自己已掉进一个似乎是很深的井里。

Done

word alignment information ??

Translation Memory

- Incrementally stores translated segments. Given a new source segment it looks for **perfect or fuzzy matches**
- Matches are ranked (100%-matches on top) and presented to the user as translation **candidates** for post-editing
- A TM can be shared among and simultaneously updated by several translators working on the same project
- TMs model the style and terminology of the customers

Translation Memory

When does it help?

- on highly repetitive, such as technical manuals
- on new versions of previously translated manuals
- when several translators are working on the same project

How does it help?

- speeds up translation process
- ensures consistency across different translators

Limitations

- number of useful matches found is generally small (5-10%)

Machine Translation

Machine translation in general decomposes the language translation process into a sequence of rule applications.

In statistical MT:

- the translation process is expressed as a search problem that computes an **optimal** sequence of rules to apply
- translation rules are automatically extracted from a **large parallel corpus** and a **stochastic model** is defined over the translation rules, that is optimised to best fit the data
- according to the employed stochastic model, the sequence of rules may generate **linear or hierarchical** structures.

Machine Translation

When does it help?

- language pairs supported by large parallel data
- translation directions between close languages
- training data represent well task data

How does it help?

- provides good draft translation to start with
- avoid translating easy/repetitive fragments

Limitations

- translations may lack of global coherence
- bad translations cause waste of time, loss of trust

TM versus MT

Capabilities	TM	MT
Can it start from scratch?	✓	
Does it improve during usage?	✓	
Can it instantly learn a new translation?	✓	
Does it consider context of the segment?		
Can it retrieve 100% matches?	✓	
Can it create new 100% matches?		✓

TM and MT are rather complementary!

Simple MT Integration

TM backed up by MT

Google code
Paid version of Google Translate

The screenshot displays the SDL Translation Management System (TMS) interface. The main window shows a translation segment for the document 'SDL Product Overview.docx' with a 100% match. The segment text is: 'When new content is written and submitted for translation SDL TMS automatically checks the content against previously translated content using the latest patented technology and advanced linguistic processing.' The right-hand pane shows a list of translation results, including a 'From Machine Translation' entry and several other entries with quality scores and references.

How to evaluate the impact of MT?

This screenshot shows a translation segment with a 100% match. The segment text is: 'The power of the solution lies in the combination of sophisticated machine translation technology with other translation automation'. The right-hand pane shows a list of translation results, including a 'From Machine Translation' entry and several other entries with quality scores and references.

Translation cost

Translation projects are quoted on word basis

Price per-word depends on:

- domain
- languages
- urgency
- quality
- TM matches

From a research perspective we are interested in the impact of MT on **user experience and productivity**

Human productivity

Daily productivity of translators is highly variable ... and also translations vary significantly among translators

To evaluate the impact of MT technology we have to consider both subjective and objective criteria:

- usability, user preferences, ...
- **variations in productivity**
 - **effort**: e.g. human TER
 - **speed**: e.g. word/hour, sec/word

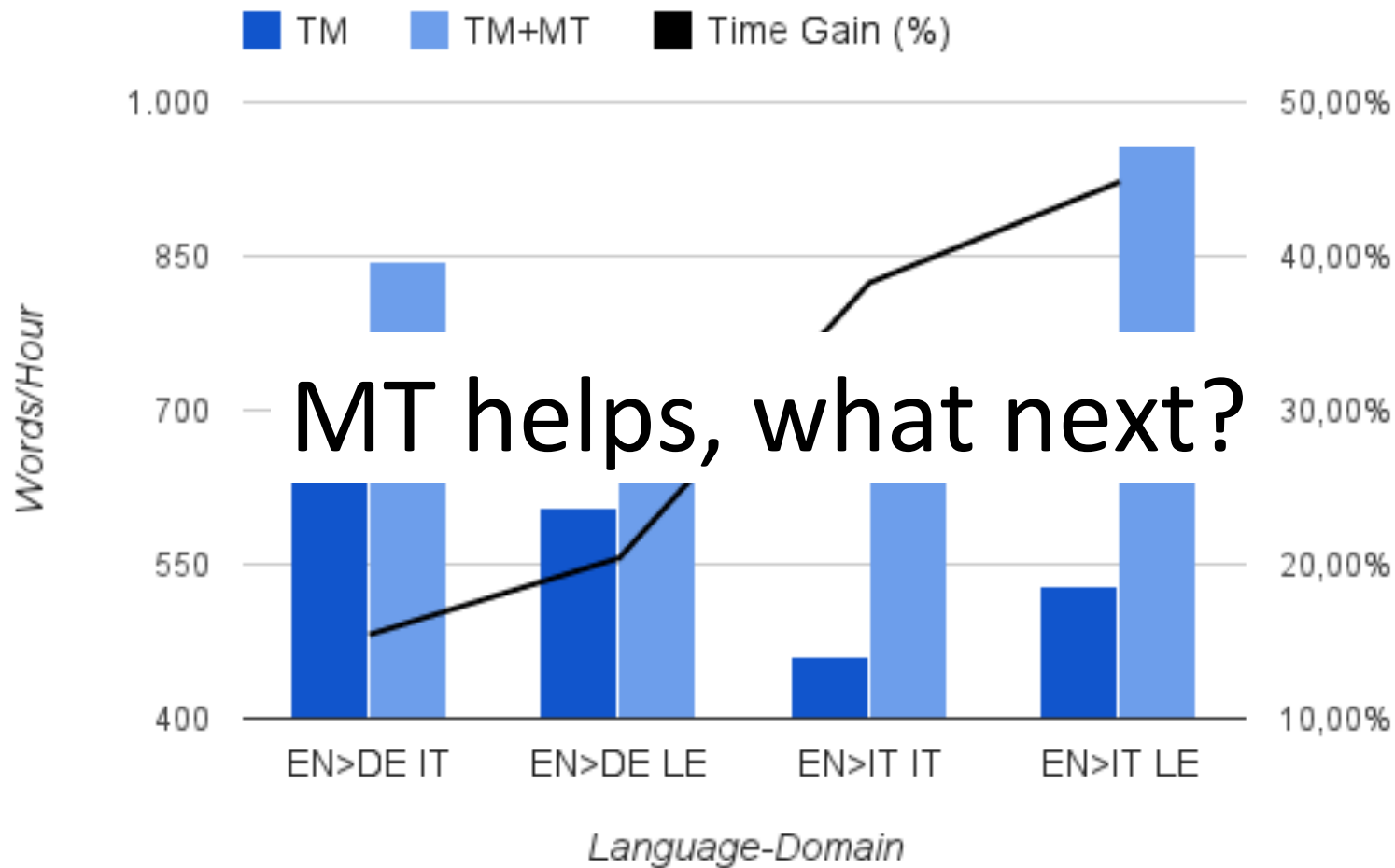
Simple MT Integration

Google code
Paid version of Google Translate



Integrated logging functions in plug-in and run experiments with 2 domains x 2 directions x 4 versions = 16 translators

Simple MT Integration





- Seamless integration of MT into the CAT workflow
- Research:
 - **self-tuning** machine translation
 - **user-adaptive** machine translation
 - **informative** machine translation
- Enterprise level open-source CAT tool
- Extensive field testing with professional translators

MateCat Tool

Translate - PROJ-359BF7E47E19CC9E903E2484864C5926 - 352

http://demo.matecat.com/translate/PROJ-359BF7E47E19CC9E903E2484864C5926/en-it/361-sldfw322d

matecat Jobs List > PROJ-359BF7E47E19CC9E903E2484864C5926 (352 - EN)

100%

A CAT Tool for Your Business. Simple. Web-Based

Un CAT Tool per il tuo lavoro. Semplice. Web-based.

T→ DRAFT TRANSLATED

Post-editing

Suggestion from TM

Suggestion from SMT

Translation matches

A CAT Tool for Your Business. Simple. Web-Based	Un CAT Tool per il tuo lavoro. Semplice. Web-based.
A CAT Tool for Your Business. Simple. Web-Based	Uno strumento CAT per il tuo business. Semplice. Web-Based
Tools for making your business competitive	Strumenti per rendere competitiva l'impresa

Source: rprosser 2006-09-19 31%

MateCat integrates Statistical Machine Translation and Collaborative Translation Memories, within the Human Translation workflow. MateCat increases the productivity of professional translators and enhances their work experience with MT.

MateCat integra la traduzione automatica statistica e le memorie di traduzione collaborativa, all'interno del flusso di lavoro di traduzione umana. MateCat aumenta la produttività dei traduttori professionisti e migliora la loro esperienza di lavoro con MT.

Progress: 31% Eq. words: -- Translated: --% Approved: --% Words last hour: -- Expected completion: -- Hours Project Statistics

Simple. Web based.

MateCat Tool

Job 984 - Editing Log
Slowest 5.000 segments by time-to-edit

Summary

Words	Avg Secs per Word	% of MT	% of TM	Total Time-to-edit	Avg PE Effort %	% of words in too SLOW edits	% of words in too FAST edits
2641	5s	96%	4%	03h:38m:59s	33%	1%	0%

Editing Details

Secs/Word	Job ID	Segment ID	Words	Suggestion source	Match percentage	Time-to-edit	Post-editing effort
163.8	984	580414	26	Machine Translation	86%	10m:58s	20%

Segment
You can move a volume to a new I/O group to balance the workload across the in the system without stopping host activity to the volumes.

Suggestion
È possibile spostare un volume di I / E in un nuovo gruppo di bilanciare il carico di lavoro tra i nel sistema host senza dover arrestare l'attività per i volumi.

Translation
È possibile spostare un volume in un nuovo gruppo I/O per bilanciare il carico di lavoro nel sistema senza arrestare l'attività dell'host sui volumi.

Diff View
È possibile spostare un volume ~~di~~ I / E in un nuovo gruppo ~~di~~ I/O per bilanciare il carico di lavoro ~~tra~~ nel sistema ~~host~~ senza ~~dover~~ arrestare l'attività ~~per~~ dell'host sui volumi.

Secs/Word	Job ID	Segment ID	Words	Suggestion source	Match percentage	Time-to-edit	Post-editing effort
28.8	984	580432	27	Machine Translation	86%	12m:57s	31%

Data collection
and logging for
in-depth analysis

```
percentage;Time-to-edit;Post-editing effort;Segment;Suggestion;Tra  
;"20%";"You can move a volume to a new I/O group to balance the  
;"31%";"Mirrored, compressed, and thin-provisioned volumes as we  
;"65%";"It is common to observe high compression ratios in datab  
;"47%";"Establish paths to I/O groups on hosts After the system  
;"46%";"Select the preferred node in that I/O group that the hos  
;"45%";"The wizard only changes the I/O group for the volume and  
8 984;580424;"Machine Translation";50;"86%";"312327";"27%";"For volumes mapped to Fibre Channel hosts, the wizard al  
9 984;580379;"Machine Translation";11;"86%";"252606";"66%";"Like thin-provisioned volumes, compressed volumes have v  
10 984;580411;"Machine Translation";37;"86%";"240454";"41%";"Planning for compression in pre-existing installations T  
11 984;580389;"Machine Translation";20;"86%";"235892";"15%";"You can also monitor information on compression usage to  
12 984;580449;"Machine Translation";22;"86%";"229242";"47%";"If a recommended service action is active, these events  
13 984;580419;"Machine Translation";28;"86%";"220055";"43%";"You can also use this wizard to move volumes to another  
14 984;580399;"Machine Translation";32;"86%";"217623";"30%";"After compression is applied to stored data, the require  
15 984;580461;"Machine Translation";29;"86%";"216299";"45%";"Email notifications The Call Home feature transmits oper  
16 984;580390;"Machine Translation";15;"86%";"198145";"44%";"To monitor system-wide compression savings and capacity,  
17 984;580465;"Machine Translation";12;"86%";"193919";"33%";"The system can send SNMP messages that notify personnel
```



- Cognitive studies of translator behaviour
 - [based on key logging and eye tracking](#)
- MT research:
 - interactive translation prediction
 - interactive editing
 - adaptive translation models
- Open source workbench
- Field test by translation agency and volunteers

Interactive MT Prediction

In May, she speaks about her return to town: "I will go there this week, but I will only stay until the 19th, because on the 20th I must present myself at work."

En mayo, habla de su regreso a la ciudad: "Voy a ir allí esta semana, pero me limitaré a quedarse hasta el siglo XIX, porque en la 20ª debo presentarme en el trabajo".

ITP T- DRAFT TRANSLATED

Translation matches

In May, she speaks about her return to town: "I will go there this week, but I will only stay until the 19th, because on the 20th I must present myself at work."

En mayo, habla de su regreso a la ciudad: "Voy a ir allí esta semana, pero me limitaré a quedarse hasta el siglo XIX, porque en la 20ª debo presentarme en el trabajo".

Source: ITP Thu Mar 2014 (v. Europe Standard Time) 75

Interactive Editing

The screenshot displays the CASMACAT web interface. At the top left is the CASMACAT logo. The top navigation bar includes 'Document list > Jobs List > tester...tester', '(5) > en > es', and a 'DOWNLOAD PROJECT' button. The main content area shows a translation of a paragraph from English to Spanish. A 'visualization >>' window is open, displaying the source text 'Nevertheless, the Tuesday announcement surprised more than one person.' and the target text 'No obstante, el martes anuncio sorprendió a más de una persona.' A blue callout bubble points to the word 'anuncio' in the target text, containing the text 'MULTIMODAL INTERACTION'. Below the translation window are 'DRAFT' and 'TRANSLATED' buttons. The bottom of the interface features a 'Translation matches' section, search filters for 'Source match' and 'Target match', and checkboxes for 'Case sensitive' and 'Regular expression'. A 'Replace' button is also present. The footer shows progress information: 'Progress: 34%', 'Total Words: 314', 'To-do: 207', 'Speed: --- Words/h', 'Completed in: ---', and a 'Reset Document' link.

Document list > Jobs List > tester...tester (5) > en > es DOWNLOAD PROJECT

58 The metropolis has had a permanent anti-corruption unit since 1870. La metrópoli ha tenido una unidad anticorrupción permanente desde 1870.

59 visualization >>

Nevertheless, the Tuesday announcement surprised more than one person.

No obstante, el martes anuncio sorprendió a más de una persona.

60 It did not need any more than that... Stephane Bergeron should call this... produced out of the hat" and as "a s... on the corner of a table"

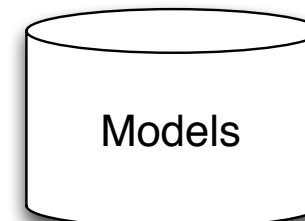
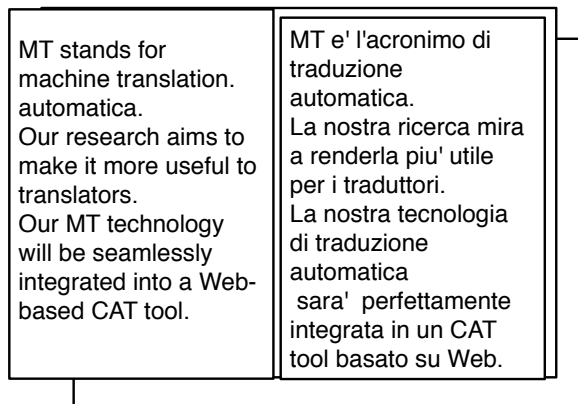
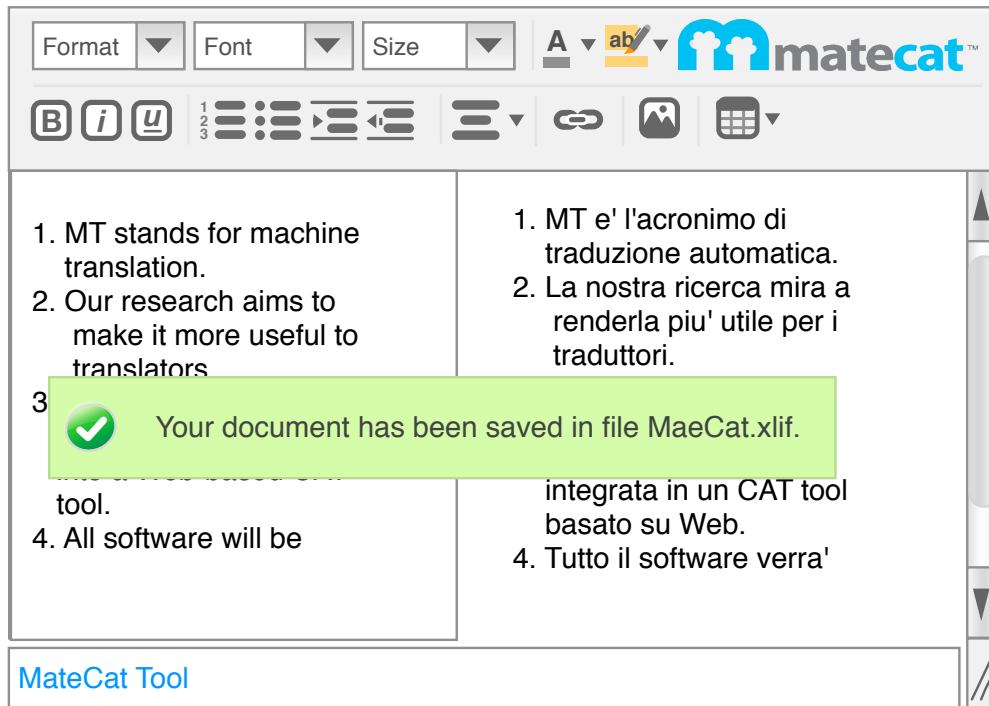
Translation matches

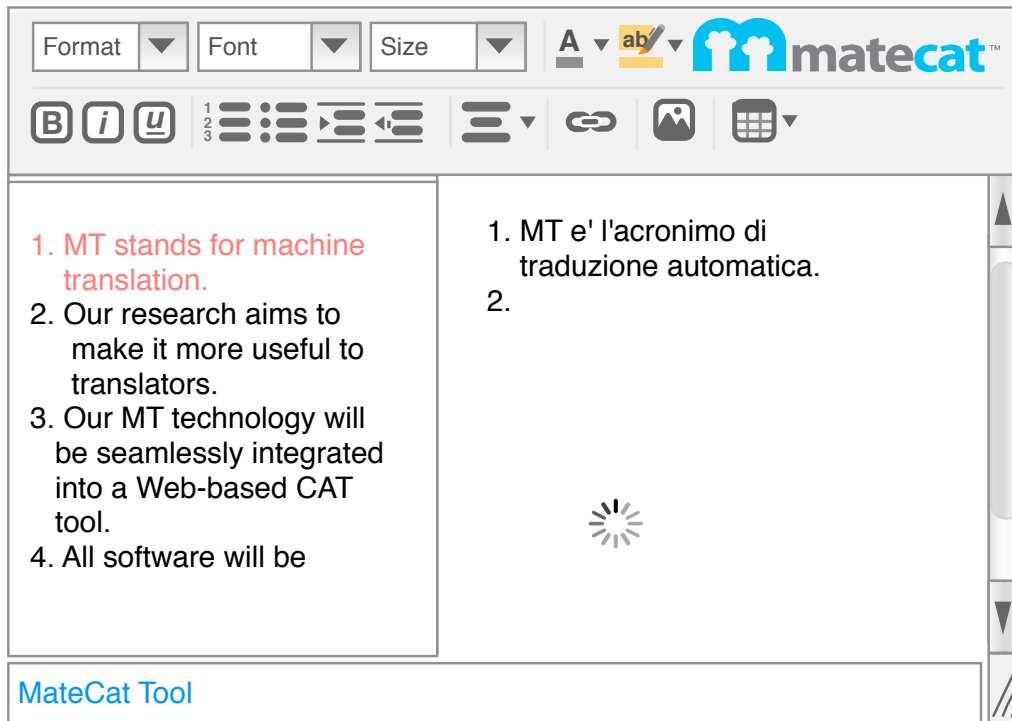
T- DRAFT TRANSLATED

Source match Target match Case sensitive Regular expression Replace

View Rules

Progress: 34% Total Words: 314 To-do: 207 Speed: --- Words/h Completed in: --- Reset Document





User-adaptive MT

User Feedback

SRC	MT stands for machine translation.
MT	MT sta per traduzione automatica.
USR	MT e' l'acronimo di traduzione automatica.

MT Server

On-Line Learning

Models

User-adaptive MT

- Discriminative re-ranking methods

K. Wäschle et. al, Generative and discriminative method for online adaptation in SMT, MT Summit 2013.


- Cache based language and translation models

N. Bertoldi et. al, Cache-based online adaptation for machine translation enhanced computer assisted translation, MT Summit 2013.

- Discriminative online adaptation of dense/sparse features

P. Mathur et al., Online learning approaches in computer assisted translation, WMT 2013.

Informative MT

 MateCat - _D1.1_-_I[...].docx_en-US_de-DE.sdlxliff (5898) > en-GB > fr-FR

ORIGINAL PREVIEW

MateCat - D1.1 - Introduction - EN.docx_[...].sdlxliff English (GB) [en-GB] > French (France) [fr-FR] Raw words: 0

7176 Introduction Introduction

7177 The goal of this work package is to develop methods and system architectures to adapt statistical machine translation (SMT) systems.

L'objectif de cet atelier de travail est de développer des méthodes et des architectures de système d'adapter les systèmes de traduction automatique statistique (TAS).


T→ TRANSLATED

Translation matches

The goal of this work package is to develop methods and system architectures to adapt statistical machine translation (SMT) systems.	L'objectif de cet atelier de travail est de développer des méthodes et des architectures de système d'adapter les systèmes de traduction automatique statistique (TAS).	Source: MT 2013-08-28 MT
Machine translation	Traduction automatique	Source: Anonymous 2012-08-26 18%
Machine header piping systems	Collecteurs/tuyauteries	Source: TRANSLATED 0000-00-00 15%

7178 We distinguish several types of adaptation:

7179 Domain adaptation:

Progress:  0% Payable Words: 331 To-do: 310 Editing Log

Informative MT

Methods and system
translation (SMT) systems.

L'objectif de cet atelier de travail est de développer des méthodes et des architectures de système d'adapter les systèmes de traduction automatique statistique (TAS).



TRANSLATED

in architectures to adapt

L'objectif de cet atelier de travail est de développer des méthodes et des architectures de système d'adapter les systèmes de traduction automatique

Traduction automatique

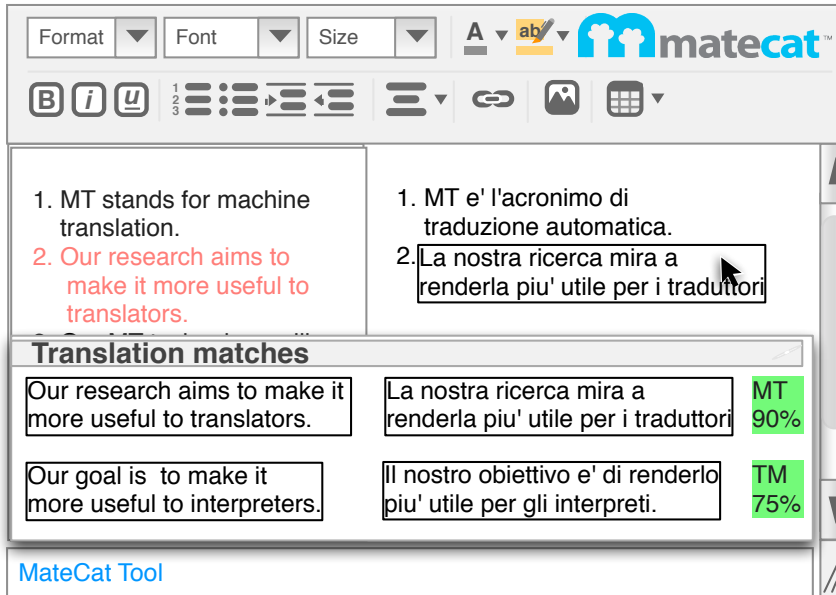
Collecteurs/tuyauteries

Source: MT	2013-08-28	MT
Source: Anonymous	2012-08-26	18%
Source: TRANSLATED	0000-00-00	15%

Informative MT

- **Score MT vs TM suggestions**
 - Show most useful suggestion in first position
- **Filter out bad MT suggestions**
 - Avoid translator wasting time, loosing trust
- **Provide reliable partial information**
 - Show suggestions of important/difficult words

Informative MT



**Informative
MT**

MT and TM suggestions

**Filtering and
ranking**

Source

SRC Our research aims to make it more useful to translators.

MT Server

MT decoder

QE engine

TM Server

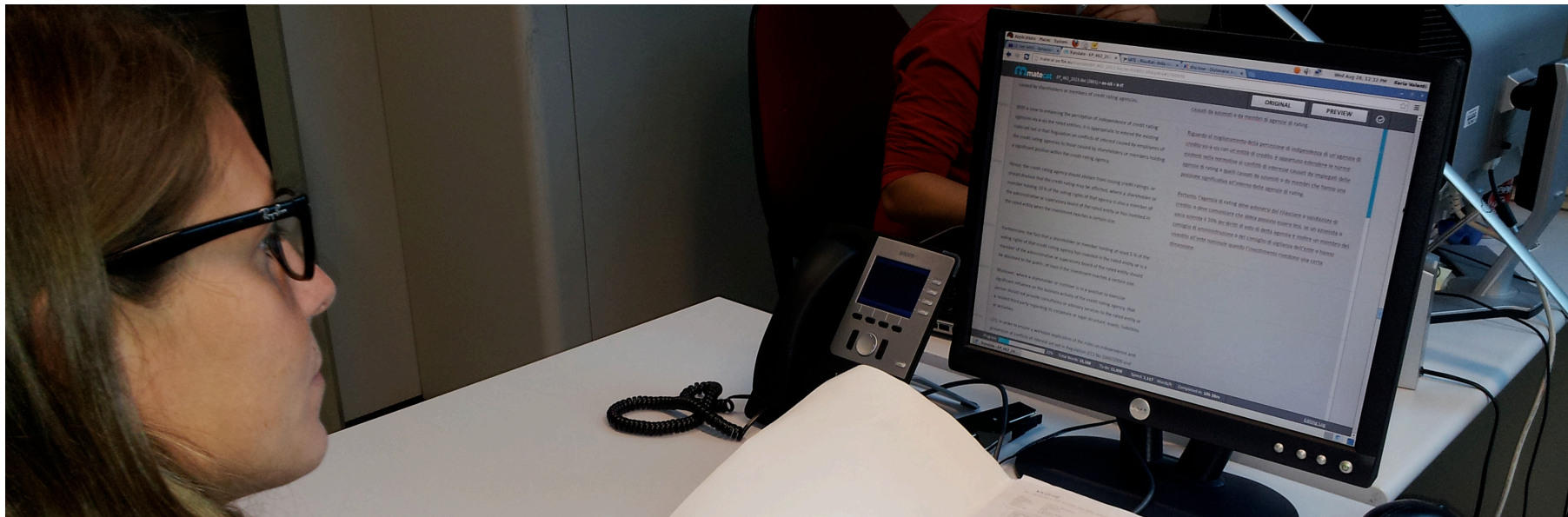
Open experimental platforms

Easy to install and run post-editing experiments

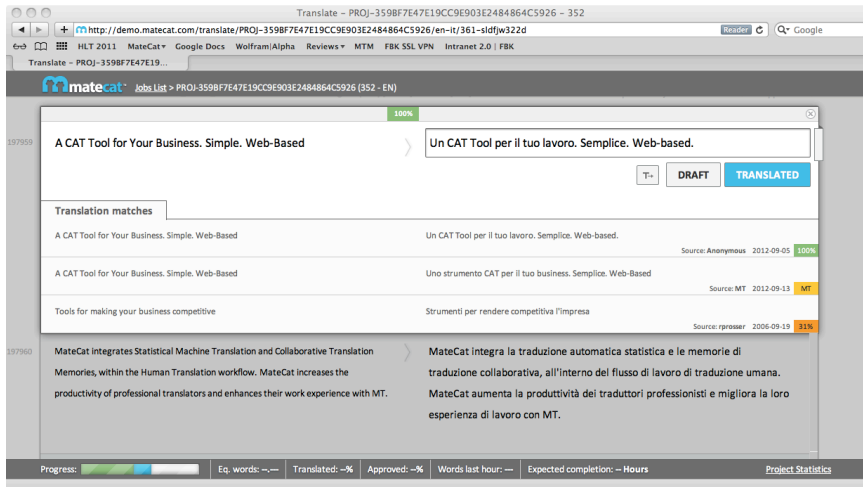
Google-compliant MT API, MyMemory TM API

comparing different MT engines, and much more

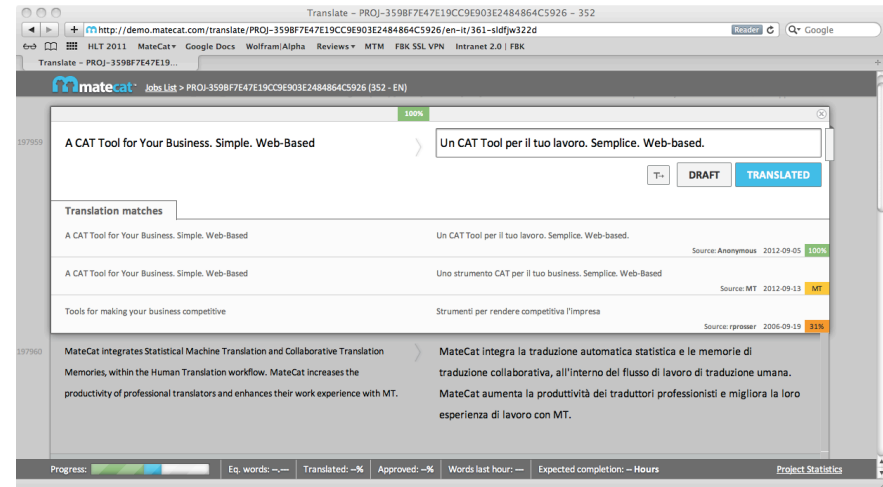
Suited for remote access (e.g. crowdsourcing)



Test Protocol

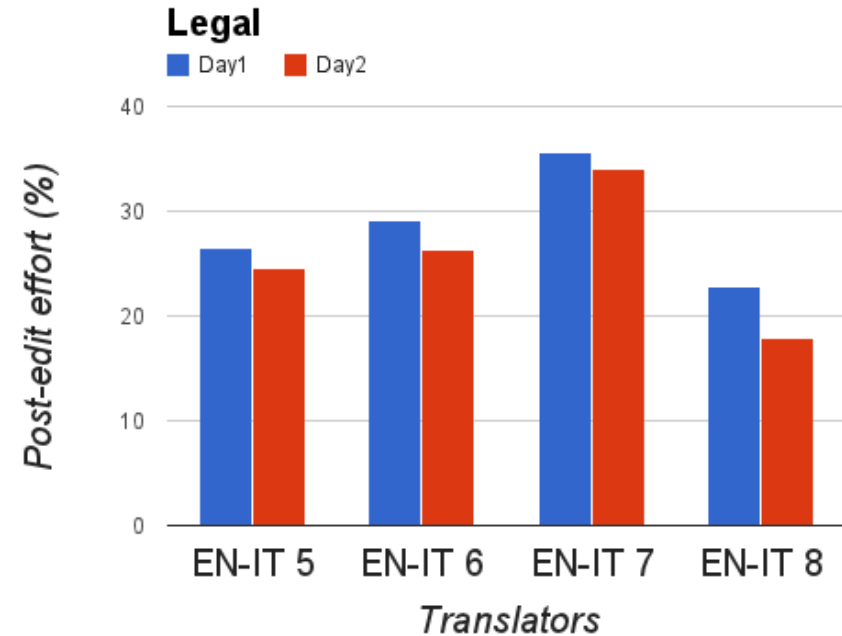
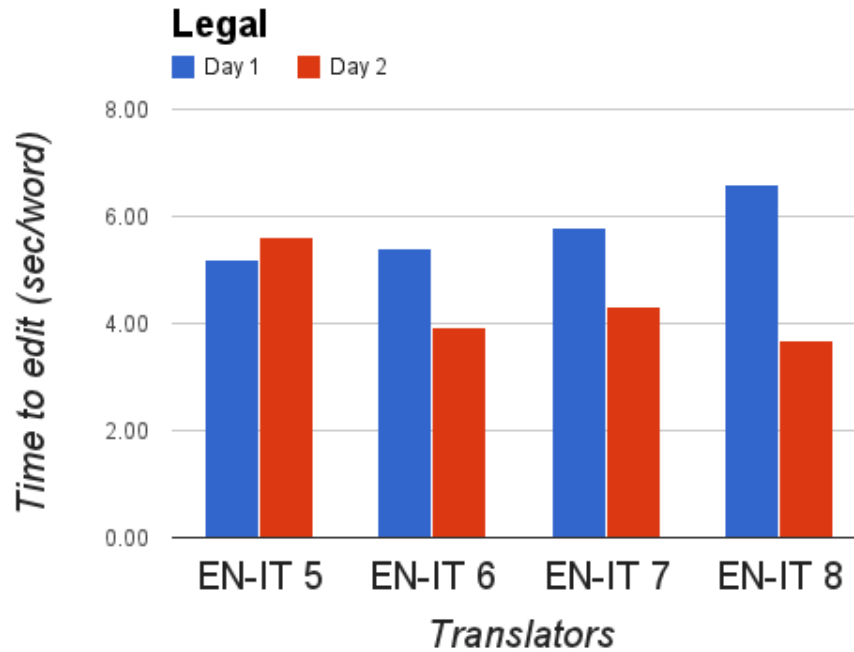


Day 1
Translation of 50% of doc
with MT1 (domain adapted)



Day 2
translation of rest of doc
with MT2 (project adapted)

Self-tuning MT

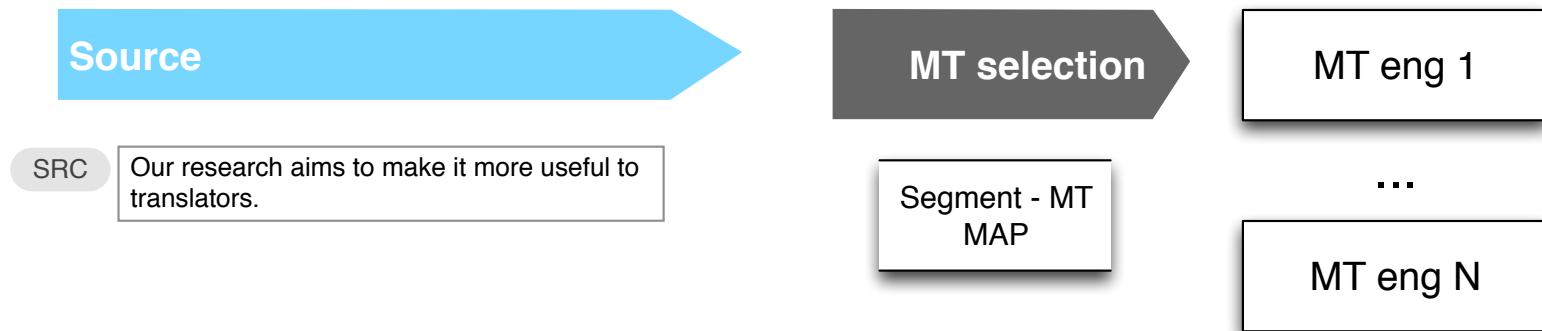
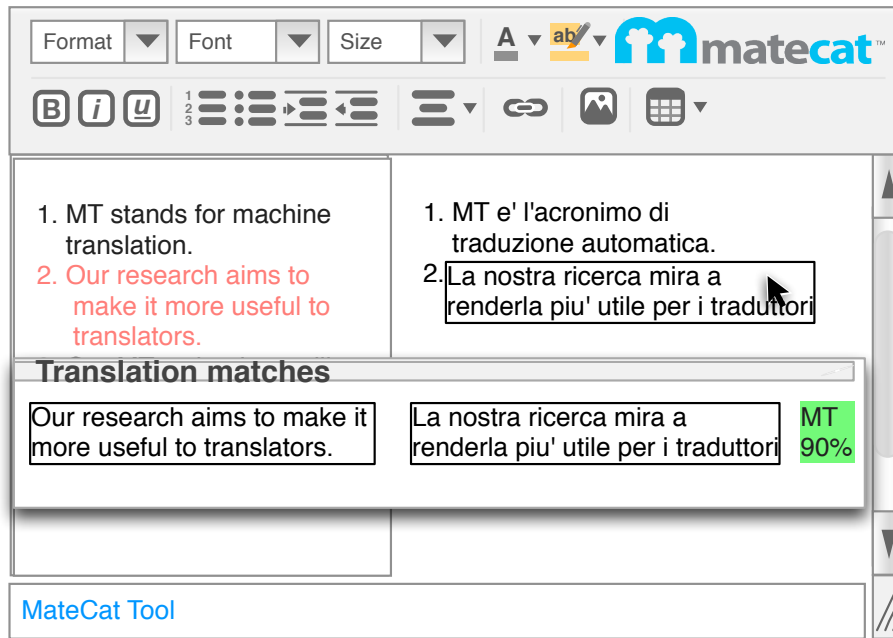


Average gain: 22.25%

Average gain: 10.71%

This protocol introduces secondary effects:
learning curve of users about system and document

Better Test Protocol



MT systems randomly switched inside the same document:
user does not know where the suggestions come from

Conclusions

- Integration of HT and MT is still an open issue
 - very challenging research problems
- Open experimental infrastructure
 - permits to evaluate how useful MT is
- **Interested to try our CAT tools?**
 - www.matecat.com simple UI, industry ready ready
 - www.casmacat.eu enhanced UI, research oriented

Thank you!

