IBM Deployment of Real Time Translation Services

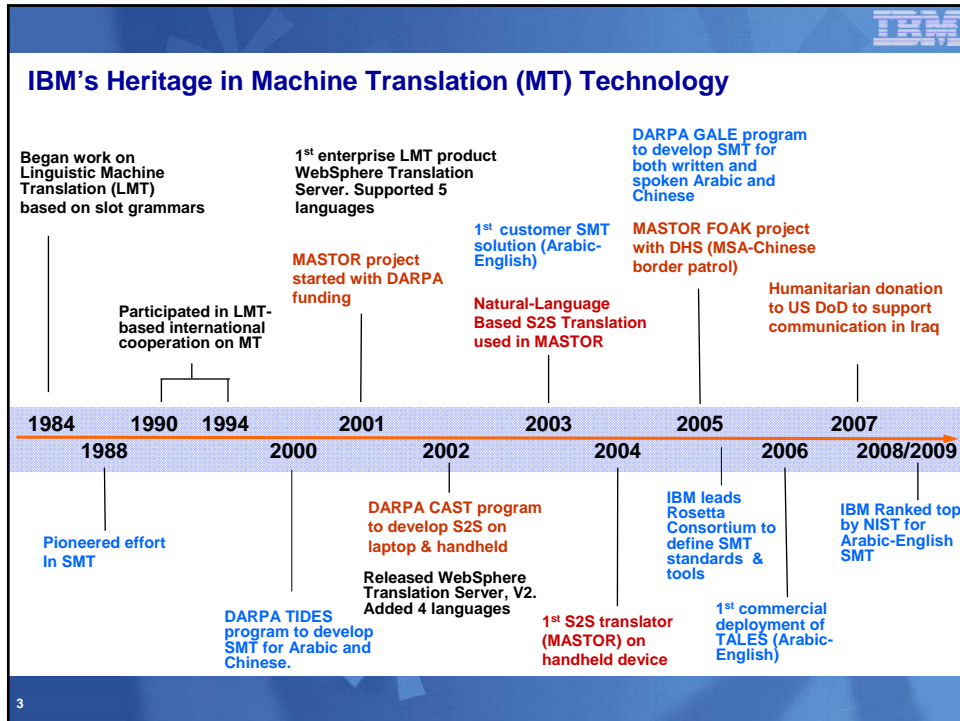**IBM T.J Watson Research Center**
David Lubensky and Salim Roukos
davidlu@us.ibm.com, roukos@us.ibm.com

Machine Translation is a Key element of the Global Enterprise

- **What are the key questions with IT customers?**
  - How do I use technology to improve service, reduce costs and manage risk in my business? In other words, how do I make my IT resources as dynamic as my business --capable of adapting to my business needs?
  - That's what CEOs are looking for and that's where the money is

- **What's needed to succeed with MT inside the Global Enterprise**
  - Committed Senior Management
  - Pervasive, enterprise-wide support
  - Secure, customizable, extensible, enterprise-strength, real-time translation
  - Technology
  - Access to Data (parallel corpora, content, etc)
  - Lifecycle management

IBM's Heritage in Machine Translation (MT) Technology

Began work on Linguistic Machine Translation (LMT) based on slot grammars

1st enterprise LMT product WebSphere Translation Server. Supported 5 languages

DARPA GALE program to develop SMT for both written and spoken Arabic and Chinese

1st customer SMT solution (Arabic-English)

MASTOR FOAK project with DHS (MSA-Chinese border patrol)

MASTOR project started with DARPA funding

Participated in LMT-based international cooperation on MT

Natural-Language Based S2S Translation used in MASTOR

Humanitarian donation to US DoD to support communication in Iraq

1984	1990	1994	2001	2003	2005	2007
1988		2000	2002	2004	2006	2008/2009

Pioneered effort In SMT

DARPA CAST program to develop S2S on laptop & handheld

IBM leads Rosetta Consortium to define SMT standards & tools

IBM Ranked top by NIST for Arabic-English SMT

Released WebSphere Translation Server, V2. Added 4 languages

DARPA TIDES program to develop SMT for Arabic and Chinese.

1st S2S translator (MASTOR) on handheld device

1st commercial deployment of TALES (Arabic-English)

3

Global Enterprise and Next Generation of Business Transformation

- IBM
  - About 400,000 employees in more than 160 countries
- Linguistic Diversity
  - Not everyone speaks the same language
    - When properly harnessed… results in incredible strength
    - When left unattended… creates huge barriers
- Transforming the Global Enterprise
  - Improve Collaboration
    - Communicate in our native languages with other colleagues (and clients) using Instant Messaging, and Mobile devices
    - Mine/share business artifacts/knowledge from documents written in other languages
    - Browse/extract content from the web and/or knowledge bases in other languages
  - Reduce the cost of translation across the corporation
    - Impact IBM spent on Localization
    - Customer Support Portals
    - User/Community created content that is NOT in anyone's budget for Translation
- Status
  - We've started the Transformation

4

n.Fluent and Real Time Translation Services (RTTS)

History

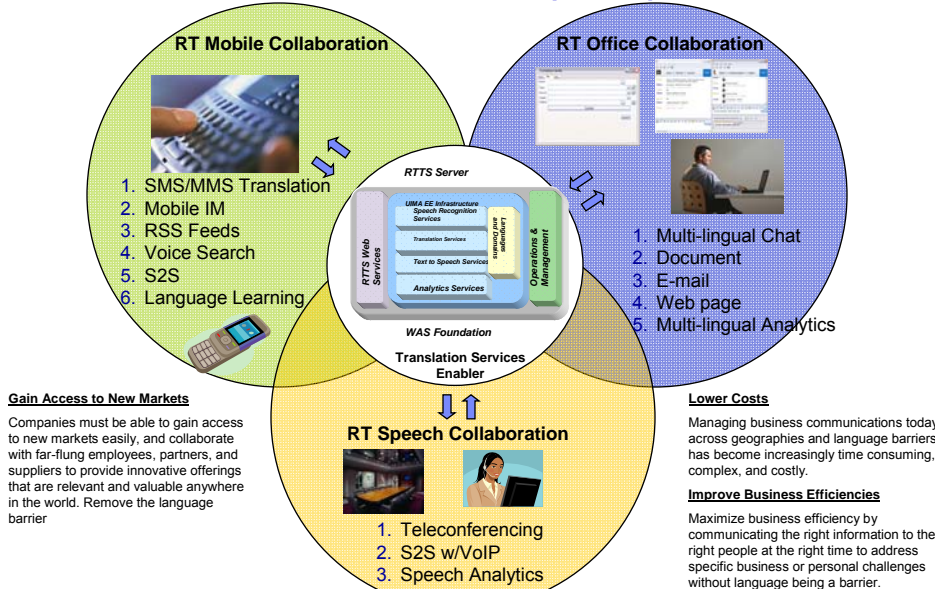2006   2007   2008   2009

innovation**jam**

- RTTS is a 2006 InnovationJam project which created a platform for speech and text translation, developed by IBM Research
- The n.Fluent project is a partnership between the CIO Office and IBM Research to create Text Translation Services, started in June 2008 and sponsored by the IBM Executive Innovation &Value Team
- The n.Fluent project team is comprised of people from around the world – Watson Research Center, China Research and Development Labs, CIO Office, Tokyo Research Lab, Haifa Research Lab
- We are collaborating with many IBM organizations – Software, Services, and CIO office

---

Real Time Translation Services (RTTS) Suite

RT Mobile Collaboration

1. SMS/MMS Translation
2. Mobile IM
3. RSS Feeds
4. Voice Search
5. S2S
6. Language Learning

RT Office Collaboration

1. Multi-lingual Chat
2. Document
3. E-mail
4. Web page
5. Multi-lingual Analytics

RTTS Server

UMA EE Infrastructure
Speech Recognition Services

RTTS Web Services

Translation Services

Text to Speech Services

Analytics Services

Operations & Management

WAS Foundation

Translation Services Enabler

RT Speech Collaboration

1. Teleconferencing
2. S2S w/VoIP
3. Speech Analytics

Gain Access to New Markets

Companies must be able to gain access to new markets easily, and collaborate with far-flung employees, partners, and suppliers to provide innovative offerings that are relevant and valuable anywhere in the world. Remove the language barrier

Lower Costs

Managing business communications today across geographies and language barriers has become increasingly time consuming, complex, and costly.

Improve Business Efficiencies

Maximize business efficiency by communicating the right information to the right people at the right time to address specific business or personal challenges without language being a barrier.

n.Fluent Overview

- Secure, enterprise-strength, real-time translation for the Global Enterprise
- n.Fluent Machine Translation Portal
  - Cut&PasteText
  - Web pages
  - Documents
  - Instant MessagingSametime chats (via a Sametime plug-in)
  - Mobile (BlackBerry and others) translation application
- Crowdsourcing – leverage 400K IBMers
- Language Pairs: English to/from Arabic, Simplified and Traditional Chinese, French, German, Japanese, Korean, Italian, Portuguese, Russian, and Spanish.

العربية   Deutsch   Français   日本語   한국어   Español

中文   English   Italiano   Português   Русский

- APIs to RTTS and n.Fluent services allow other applications to access speech and translation services
- In the process of moving from the Technology Adaption Program (TAP) environment to a 24x7 supported environment

---
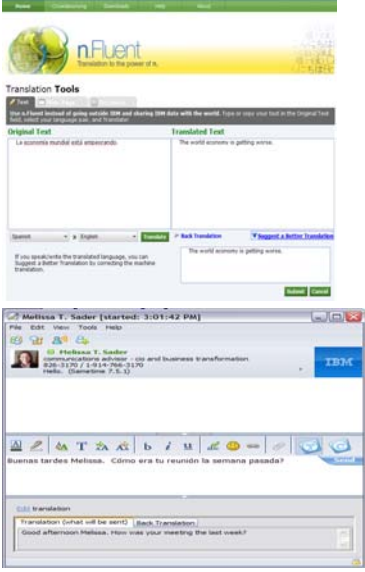
n.Fluent Translation System
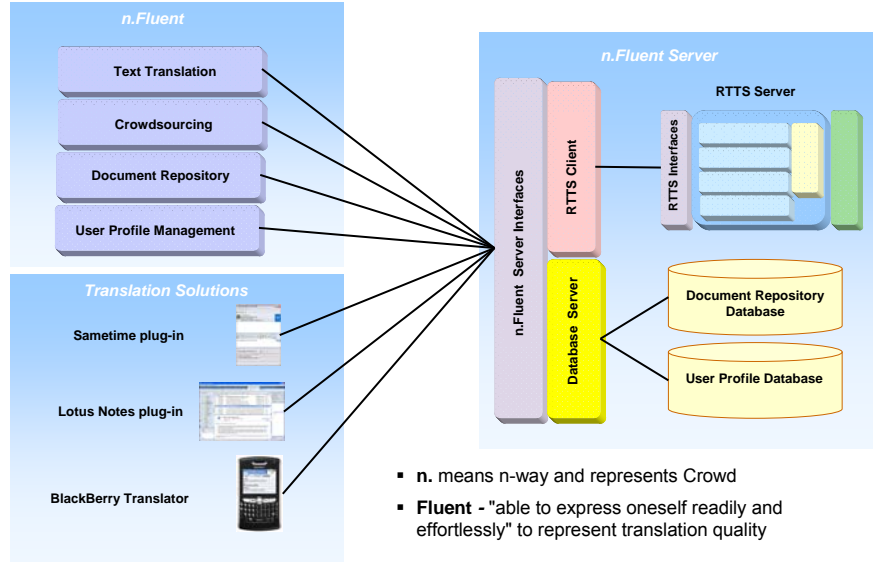
- **n.** means n-way and represents Crowd
- **Fluent -** "able to express oneself readily and effortlessly" to represent translation quality

n.Fluent Text Translation Portal

---

n.Fluent Activity / Usage

Activity	Current Totals
Registered Participants	2,720 users
Crowdsourcing (chat contributions)	4.2 M words
Crowdsourcing (translation)	125K words
Webpage translations	70 K
Number of words translated	27.5 M words
Downloads (IM plug-in, Mobile Translator)	2,900
Internal Deployments	8
Projects (in process)	6

Internal Deployments
w3 News
IT Help Central
ThinkPlace
SWG – Globalization
SWG – Learning for Growth
GBS Incentives Workplace
ITS – Integrated MBP Delivery
CHQ WW Expense Reporting

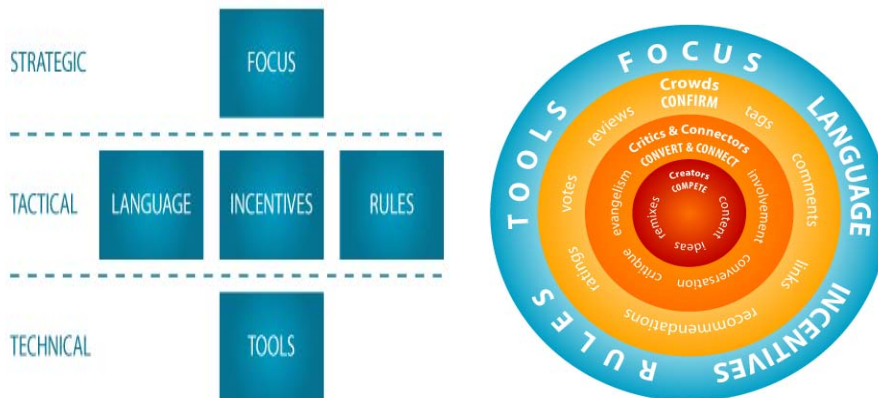| Current Projects |
| --- |
| IBM Support Portal |
| SWG WPLC wikis |
| TSC Post-editing with MT |
| GBS – Contribute Asset |
| Self Enablement Portal |
| www.ibm.com |

n.Fluent Activity – Words Translated by Language

Language	Text	Web Page	Document	Total
Arabic	264,670	87,280	511,695	863,645
Chinese - Simp.	268,424	326,814	85,895	681,133
Chinese - Trad.	116,602	0	14,077	130,679
French	2,559,907	2,442,525	448,969	5,451,401
German	1,773,058	1,645,675	3,413,085	6,831,818
Italian	1,303,723	1,053,644	1,545,856	3,903,223
Japanese	627,881	146,803	50,733	825,417
Korean	174,763	16,709	38,739	230,211
Portuguese	1,332,574	875,463	94,292	2,302,329
Russian	162,830	15,236	21,843	199,909
Spanish	2,153,684	2,233,352	1,034,423	5,421,459
Totals	10,738,116	8,843,501	7,259,607	26,841,224

Building a Crowdsourcing Community

The FLIRT model of Crowdsourcing (Sam Viitamäki 2007):

FLIRT = **F**ocus (align business needs/users); **L**anguage (Social objects/attracting); **I**ncentive (motivating/retaining); **R**ules (interacting); **T**ools (platform)

n.Fluent
Translation to the power of n.

TRANSLATION CHALLENGE
The Million Word Quest

News
Top stories Past 7 days >

Crowdsourcing – Corporate Communication is key

- **Major Boost in participation - Mind Blowers w3 News article**
  - First in w3 news series that highlights IBM's innovation – 16K page views
  - n.Fluent new registrations = 949

- **What is the n.Fluent Language Translation Challenge?**
  - It's a worldwide, internal, online Crowdsourcing event in July'09
  - Goal is to drive Crowdsourcing volume to use as input to translation improvements
  - In return, top individual and language group contributors will be recognized with monetary donations to a charitable organization

Mind Blowers:
n.Fluent

Episode 1: IBMers break down language barriers with automated translation. [Profiled for all IBM]

13

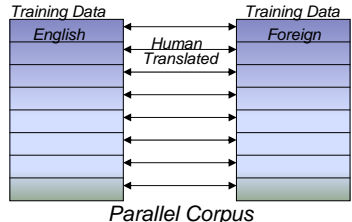SMT-Based Technologies Increases Accuracy through 'Domains'

Domain: Topic of conversation that the translation engine will recognize and translate

Custom domains are created by enhancing a broad base domain, and must be created for each language pair

Domain Examples:

Military	Custom	IT Help
Travel	Medical	Commerce
Conversational Adaptations		
Base Domains (News)		

An SMT domain model is based on a 'parallel corpus'

Training Data
English

Human Translated

Training Data
Foreign

Parallel Corpus

Domain Model creation is an iterative process

Collect
Clean
Translate
Build
Test

Domain Model Creation Steps
1. Collect training data in both languages
2. Review and 'clean' the data
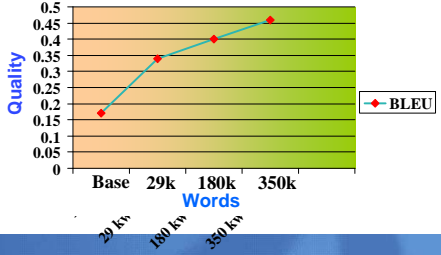3. Translate each sentence to the other language to create the parallel corpus
4. Execute the model building tools to create the statistical models
5. Test the created models
6. Repeat 1-5 refining models and improving weakness identified during accuracy testing
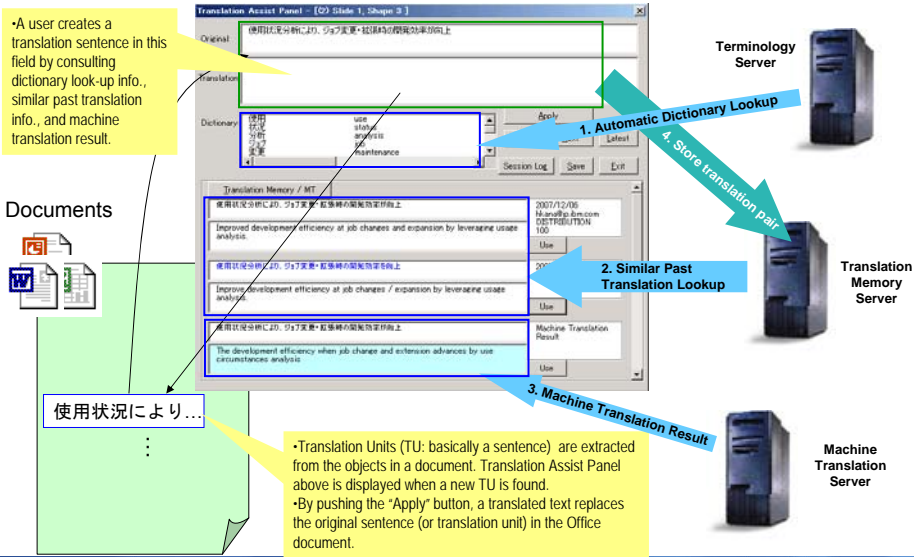
The quality / size of the domain model directly impacts the translation quality

Quality

0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0

Base   29k   180k   350k
Words

29 kw   180 kw   350 kw

◆ BLEU

14

Translation Assist Tool (TAT)

•A user creates a translation sentence in this field by consulting dictionary look-up info., similar past translation info., and machine translation result.

Documents

1. Automatic Dictionary Lookup

4. Store translation pair

2. Similar Past Translation Lookup

3. Machine Translation Result

Terminology Server

Translation Memory Server

Machine Translation Server

使用状況により…

•Translation Units (TU: basically a sentence) are extracted from the objects in a document. Translation Assist Panel above is displayed when a new TU is found.
•By pushing the "Apply" button, a translated text replaces the original sentence (or translation unit) in the Office document.

Next steps/Roadmap

- **Continue to deploy n.Fluent across the Global Enterprise**
  - Customer Support Portal (both internal and external content)
  - Localization
  - Enable Translation on the Internal websites (100M internal webpages)
  - Leverage Translation as part of multilingual analytics (Smarter Planet)
- **Continue to drive improvements in MT – leveraging the multilingual workforce to contribute conversational data via Instant Messaging and crowdsourcing**
- **Streamline vertical customization process**
- **Develop additional languages**