MT Customization

Rémi Zajac
Director of Computational Linguistics
SYSTRAN Software, San Diego

Content

1. Why MT customization?
2. Assessing a customization project
3. Overview of the customization process
4. Corpus processing and term extraction
5. Dictionary customization
6. Linguistic Customization
7. Evaluation
8. Conclusion: automating the customization process

(1) Why MT?

- Human translation
  - Not available
  - Too expensive
  - Too slow
- Polished translation not necessary
  - Gisting
  - Summary
  - Drafting

(1) Applications of MT

- Assimilation: gisting
  - Intelligence analysis
  - WWW
- Dissemination
  - Post-edited: translation/localization; TMs
  - No post-editing:
    - Technical: maintenance, support,… (Cisco)
    - Non-technical: administrative (EU)

(1) Why MT Customization?

- Higher quality:
  - Provide raw MT for large knowledge bases
  - Minimize post-editing for technical documents
- Precision and consistency in technical term translation over large amount of documents
- User maintained dictionaries

(1) When to use customized MT

- Large document knowledge base:
  - Over 10,000 pages
- Emphasis on content accessibility over polished language, but
  - Precise translation is required
  - Documents must be readable
  - Document contents must be understandable

(1) Cost/benefits analysis: CRM

- CRM Average Cost per Transaction (Forrester, 2001):
  - Telephone: $32.74
  - Web self-service: $1.17
- Web site:
  - 10,000 hits/day, 10% less calls: save $30,000/day
  - A $1M customization effort is recouped in 1 month

Content

1. Why MT customization?
2. **Assessing a customization project**
3. Overview of the customization process
4. Corpus processing and term extraction
5. Dictionary customization
6. Linguistic Customization
7. Evaluation
8. Conclusion: automating the customization process

(2) Translatability and Linguistic Closure

- Translatability: how close is the natural language used in the domain to the artificial language implemented in the MT system?
- Linguistic closure: some domains and document types use very repetitive and uniform language: few new words or syntactic constructions appear after having seen a few hundred documents.

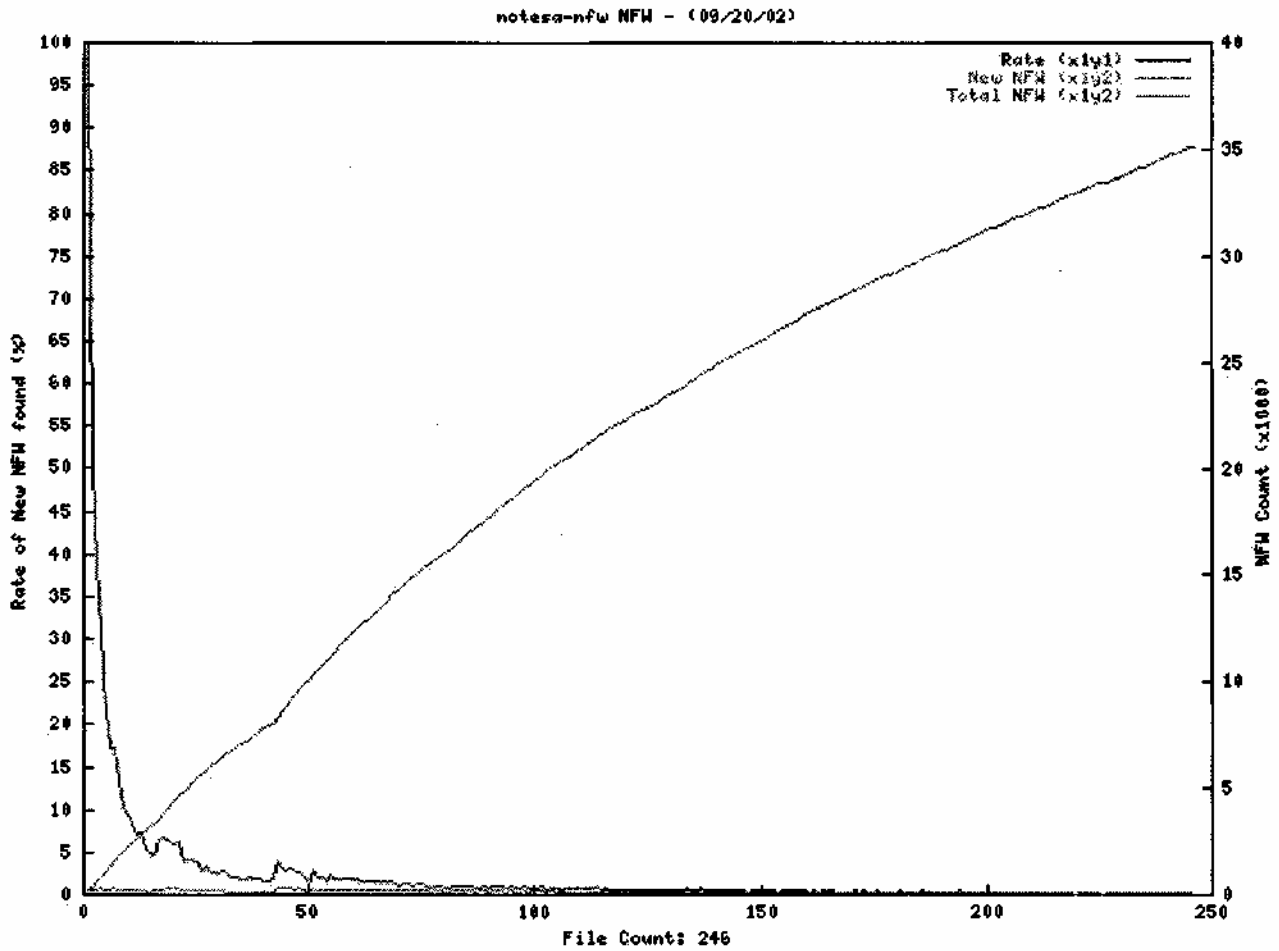
(2) Measuring Translatability

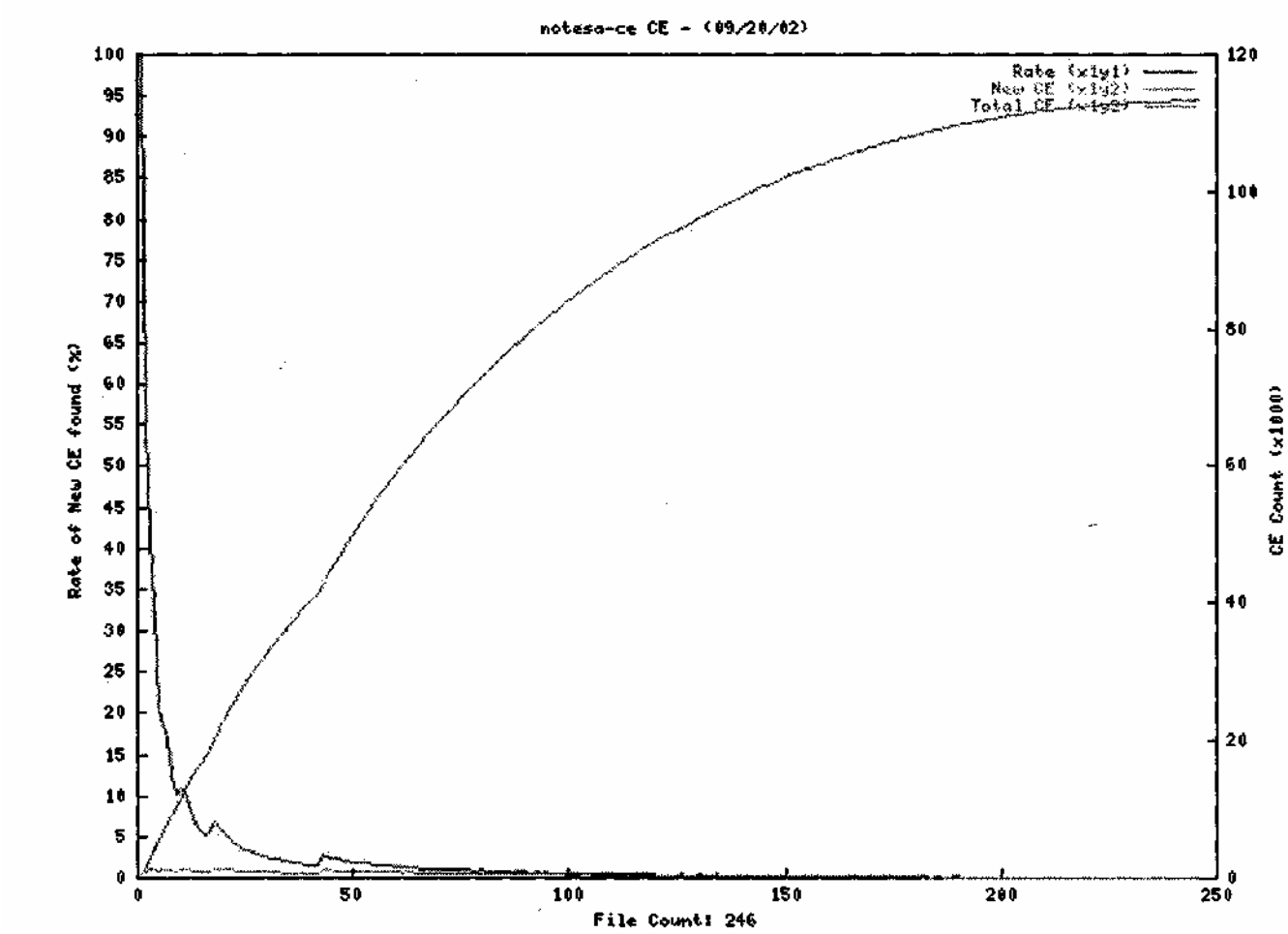
- Eg, IBM translatability index (Bernth99)
- Frequency of
  - -ing words
  - Problematic grammatical words ('to',...)
  - Complex coordinate structures
  - Complex sub-clauses
- Suppressing problematic constructions

  → Controlled Languages

(2) Linguistic Closure

- Lexical closure:
  - Count open-class words/expressions not in the system's dictionaries
  - Check lexical growth graphs
- Syntactic closure:
  - Approximate by counting function words
- Idiomaticity:
  - Check for flowery language, metaphors, culturally loaded constructions.


notesa-nfw NFW - (09/20/02)
Rate (x1y1)
New NFW (x1y2)
Total NFW (x1y2)
Rate of New NFW found (%)
NFW Count (x1000)
File Count: 246
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
0
5
10
15
20
25
30
35
40
0
50
100
150
200
250


notesa-ce CE - (09/20/02)
Rate (x1y1)
New CE (x1y2)
Total CE (x1y2)
Rate of New CE found (%)
CE Count (x1000)
File Count: 246
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
0
20
40
60
80
100
120
0
50
100
150
200
250

(2) Customization assessment

- Start from an existing system
- Measure gap between current and target translation quality
- Typical unbridgeable gap between target quality (human-level) and achievable quality: 10-20%

(2) Customization assessment

1. Measure lexical gap:
    1. Extract terms not in the systems dictionaries
2. Measure syntactic gap:
    1. Select representative documents and build custom dictionary
    2. Manually assess number and frequency of translation problems (Analysis, Transfer, Generation)
3. Identify unbridgeable issues:
    1. Format: no format or poorly formatted documents
    2. Spelling: spe!ling errors, abbreviations, etc.
    3. Language: poor grammar, telegraphic style, punctuation, etc.
4. Assess level of achievable quality (and effort)

    → This produce initial customization plan

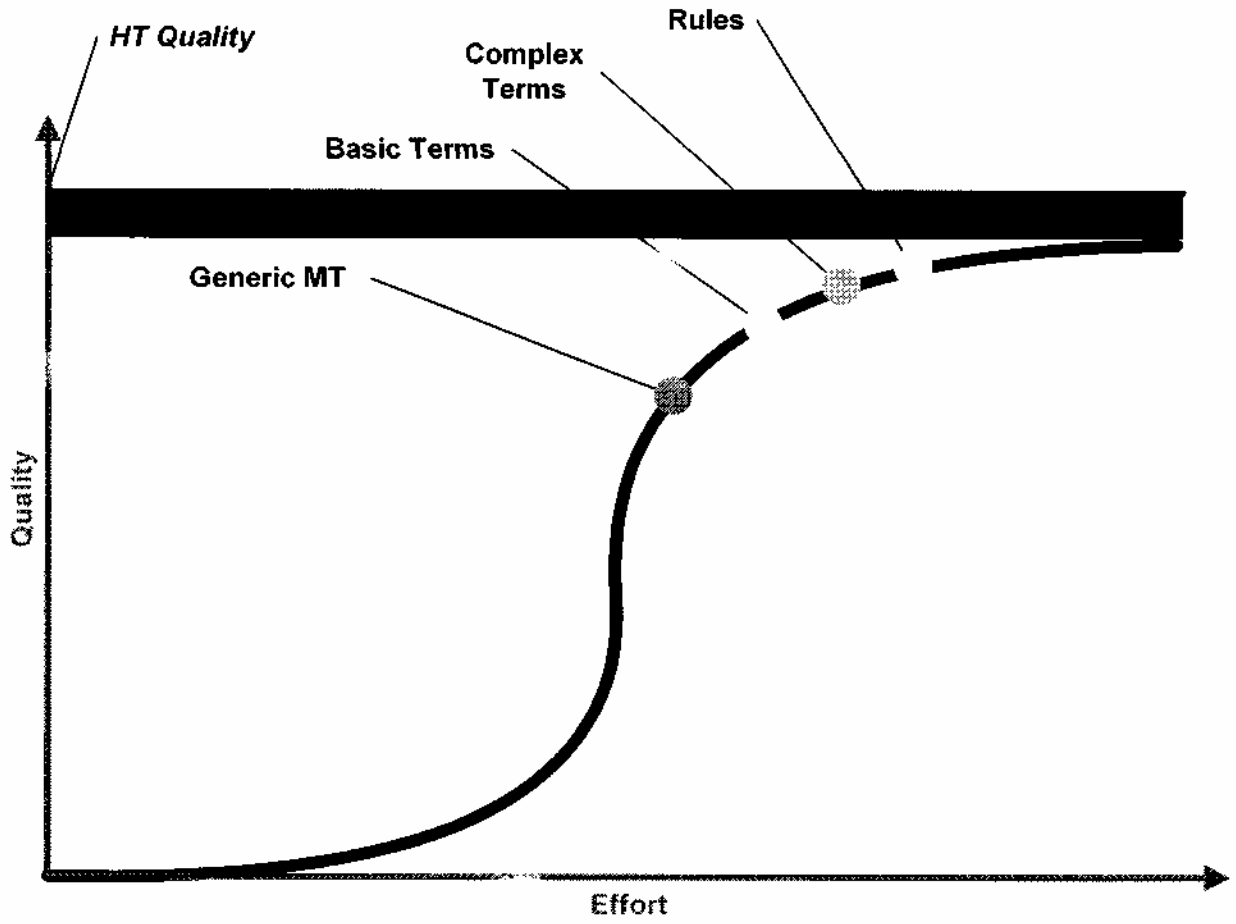
Content

1. Why MT customization?
2. Assessing a customization project
3. **Overview of the customization process**
4. Corpus processing and term extraction
5. Dictionary customization
6. Linguistic Customization
7. Evaluation
8. Conclusion: automating the customization process

(3) Customization process outline

1. Corpus processing and term extraction
2. Term translation and coding
3. Assess linguistic customization needs
4. Implement customized linguistic rules
5. Evaluate customized system
6. Fine-tune dictionary and rules
7. User testing

(3) Customization effort



Content

1. Why MT customization?
2. Assessing a customization project
3. Overview of the customization process
4. **Corpus processing and term extraction**
5. Dictionary customization
6. Linguistic Customization
7. Evaluation
8. Conclusion: automating the customization process

(4) Term extraction

- Words not in the system's dictionaries
  - 'Not Found Words' - NFWs
- Expressions that should not be translated
  - 'Do Not Translate' - DNTs
  - proper names, product names, command names, etc.
- Multiword expressions not in the system's dictionaries
- Most frequent words/expressions found in dictionary
- Subcategorization patterns: verbs, deverbals, adjectives, adverbs.

(4) NFWs

4419	Openvms	1021	Executables	663	apps	503	jserv
2843	Metalink	967	Exec	635	adpatch	495	sysdate
2739	Patchset	905	Readme	634	netware	482	multibyte
2614	Flexfield	869	Dvoid	629	intermedia	446	dbkey
2258	Sqlplus	852	metadata	624	e-business	442	subinventory
1713	webdb	800	ename	618	ias	429	unhandled
1525	jinitiator	783	flexfields	601	datatypes	426	appletviewer
1454	controlfile	780	autoinvoice	593	setenv	422	desc
1345	hp-ux	780	sid	574	istore	411	loadjava
1329	datatype	755	logminer	562	relinking	390	jserver
1243	emp	739	rowid	558	multi-org	387	jdeveloper
1212	bugcheck	726	cd-rom	543	rel	383	iprocurement
1179	relink	719	deptno	511	adadmin	382	multiversion
1123	mini-pack	707	empno	509	minipack	380	bugchecks
1113	servlet	668	compaq	506	archivelog	373	characterset

(4) DNTs

Proper Name	**Association of Support Professionals**
Acronyms	**WWW**
Software commands	**Click Yes**
Product names	**Microsoft Word**
File names	**http://www.systrans.com**
Misc.	**best effort**

(4) MWEs

3865	windows nt	1624	concurrent manager	1189	enterprise edition
3477	operating system	1580	standby database	1184	environment variables
3195	error message	1495	enterprise manager	1177	package body
2605	solution description	1483	primary key	1157	web server
2334	java class	1471	control file	1140	installation guide
2238	environment variable	1459	applications release	1109	logical name
2186	software errors fixed	1412	parallel server	1088	http server
2164	command line	1378	exec sql	1088	sales order
2110	application server	1378	database link	1036	general ledger
2046	known problems	1334	family pack	1035	configuration file
1948	rollback segment	1251	rdb release	1034	file name
1940	sql statement	1226	system administrator	1003	concurrent program
1825	release notes	1224	intelligent agent	981	user's guide
1824	storage area	1217	default value	941	file system
1765	search words	1195	error messages	904	storage map

(4) Frequent Words/Expressions

abandon	verb	If no carrier is detected within the specified time, the call is abandoned.
abort	verb	Aborted frames.
abort	noun	22 input errors, 0 CRC, 0 frame, 0 overrun, 22 ignored, 0 abort.
absence	noun	Absence of downstream digital modulated signal.
access	noun	The device provides physical layer T1 access.
access	verb	How do I access the Netscape FastTrack administrative server?
account	noun	Will it allow dynamic creation of accounts?
accounting	noun	Accounting.
achieve	verb	Achieve Optimal Routing.
action	noun	That action can cause poor performance.
activate	verb	During a switchover, the secondary protocol activates the local interface.
active	adj	What command displays the active console?
active	noun	commands are " defined and active.
addition	noun	The interface command requires additions.
address	verb	XOPR addresses the need for network connections.
address	noun	Why can not I ping my own interface address?
adjustment	noun	The modem continues to transmit requests and perform adjustments.

(4) Lexical Patterns

Relationship	*Extracted instance*	*From sentence*
Verb-Object	configure <bridging>	How do I configure bridging on ARM ?
Verb-Object-Preposition	specify <direction> (in)	The direction must be specified in later software releases.
Verb-Object-Infinitival	configure <client> <obtain>	...the client is configured to obtain an IP address
Verb-Particle-Object	find out <number>	How do I find out the number of files that a process has open?
Verb-Preposition-Object	refer (to <code>)	For more details refer to the debug codes.
Noun-Preposition-Noun	configuration (for <authentication>)	Configurations for login authentication.
Adjective-Preposition-Noun	available (to <customer>)	available to end users and customers
Adjective-Preposition	equivalent (to)	is equivalent to:

Content

1. Why MT customization?
2. Assessing a customization project
3. Overview of the customization process
4. Corpus processing and term extraction
5. **Dictionary customization**
6. Linguistic Customization
7. Evaluation
8. Conclusion: automating the customization process

(5) Term translation

- Out-sourced to translation company
- Information:
  - Source term
  - POS
  - Default translation
  - Examples of sentences
  - (Examples of translations)

(5) Linguistic coding

- Add linguistic codes:
  - POS
  - Inflectional class
  - Irregular stems and forms
  - Headword of expressions
- Use automated coding tools
  - Guess POS and inflectional classes
  - Guess headwords

(5) Test and debug

- Compile inflected dictionary
- Check sample inflected entries
- Run translation on test DB
- Check changed translations
- Tool show which entries were fired

Content

1. Why MT customization?
2. Assessing a customization project
3. Overview of the customization process
4. Corpus processing and term extraction
5. Dictionary customization
6. **Linguistic Customization**
7. Evaluation
8. Conclusion: automating the customization process

(6) Major linguistic issues

- Homographs: disambiguation rules
- PP attachment
- Conjunctions and complex clauses
- Transfer rules for function words

(6) Homographs

Term	*Guessed*	*Correct*
(go to) feature	**Verb**	**Noun**
drop	**Noun**	**Verb**
console	**Verb**	**Noun**
separate	**Verb**	**Noun**

(6) PP attachment

Displays **[errors during] Phase 2**

Enable **[failover using] the following command**

Compare **the [values before] the download**

Configuring **[timers after] failed dial attempts**

(6) Transfer rules

- Check whether the errors appear
  - Subcategorized clause for 'check' should be generated as sentence in Korean
- Both of the modems work well
  - Two meanings in Korean: both sides, or both of them.

(6) Complex constructions

- Consequence:
  - Temporal
    - Install the nut and then attach the bolt
  - Causal
    - If the belt breaks, then a new one must be installed
- Obligation:
  - Factual
    - The paths must be the same.
  - Moral
    - The user must enter the password

Content

1. Why MT customization?
2. Assessing a customization project
3. Overview of the customization process
4. Corpus processing and term extraction
5. Dictionary customization
6. Linguistic Customization
7. **Evaluation**
8. Conclusion: automating the customization process

(7) Testing and evaluation

Testing

- Development-oriented
- Linguistic criteria
- Precise metrics
- Feedback to dev team

Evaluation

- User-oriented criteria
- Translation quality
- Translation usability
  - Readability
  - Utility

(7) Testing

- Regression testing
  - Database of test sentences
  - Record of previous translations
  - Evaluate sentences which translation has changed:
    - Better, Worse, Similar
  - Decide if the ratio of improvements versus degradation is acceptable

- Linguistic testing
  - Database of sentences
  - Mark translation errors (J2450):
    - Terminology
    - Grammar
    - Format
  - Score:
    $1 - (3.5*T + 3*G + 2*F)/W$
  - Generate feedback: new pbs not identified during initial term extraction and linguistic assessment

(7) What to evaluate

- Linguistic quality
  - Question: rate the quality of this translation on a scale of 0-4.
  - Negative focus: on translation errors
  - Precise metric better done using detailed linguistic testing
- Utility:
  - Question: Can you solve your problem using this document?
  - Positive focus: problem-solving
- Usability and user satisfaction
  - Positive focus: overall impression
  - Question on readability, not quality: rate the readability of this document on a scale of 0-4
- Sometimes large differences between positive and negative approaches to evaluation

(7) How (not) to evaluate

- Absolute ratings have little significance
- Need comparison points
  - Other MT systems
  - Other language pairs
  - Previous versions
- Human translation baseline mandatory

(7) Machines against humans

- Human translation is the benchmark
- But even the benchmark is not perfect:
  - Human translation typically degrades text quality
  - Can sometime improve the original if the original is badly written
- And even the original text is imperfect
  - Bad spelling
  - Incorrect grammar
  - Simply badly written and not understandable

(7) Machines against humans II

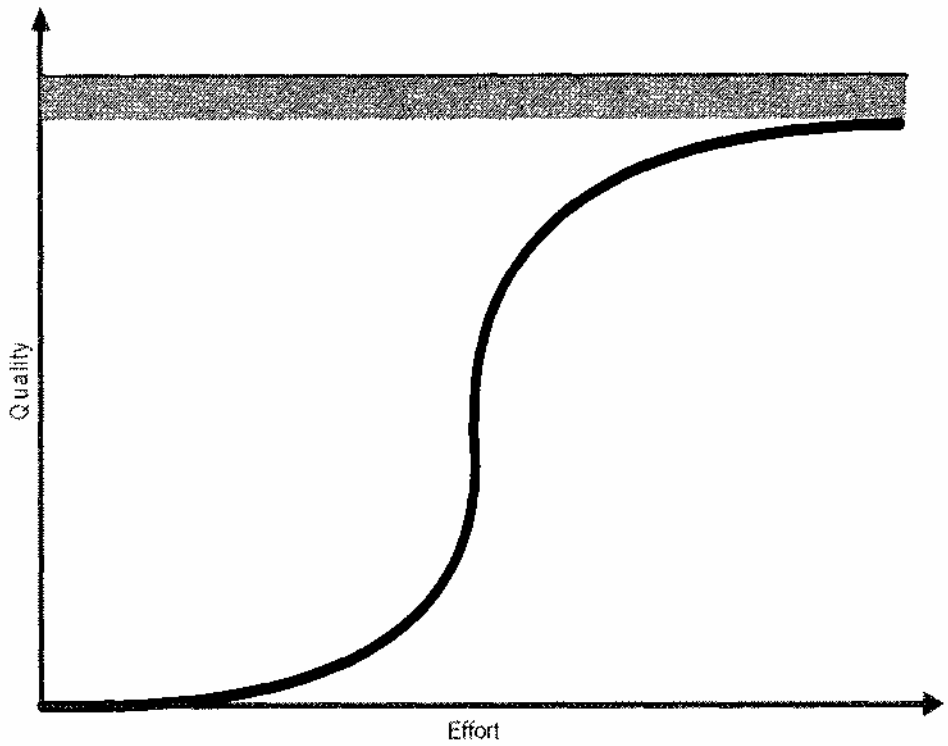
- Evaluation scheme.
  - double-blind evaluation
  - of a statistically significant set of documents
- Evaluate a mixed pool of human and machine translated documents
- Set goals as a percentage of human baseline scores

(7) Machines against humans III

- Utility scores
  - Can you solve your problem using this document?
  - Human translation: 90% of documents rated Yes
  - Machine translation: 70-95% of HT
- Readability scores
  - Rate readability of this document on a scale of 0-4
  - Human translation: 2.5-3.5
  - Machine translation: 60-90% of HT

(7) Machines against humans IV

- Early efforts bring a lot of improvements
- Later efforts follow the law of diminishing returns
- Quality gap can be narrowed, not by investing more on MT, but on controlled structured authoring



Content

1. Why MT customization?
2. Assessing a customization project
3. Overview of the customization process
4. Corpus processing and term extraction
5. Dictionary customization
6. Linguistic Customization
7. Evaluation
8. **Conclusion: automating the customization process**

(8) Automating customization

- Term extraction: improving precision/recall
- Term translation: using TM and bitexts
- Term coding: using Intuitive Coding
- Testing and evaluation: using SRM
- Linguistic assessment: SRM
- Linguistic customization: open

(8) Manual/automatic customization

- Comparison with MSR:
  - Same basic transfer architecture
  - Same level of analysis ('deep syntax')
  - Uses trained statistical parser instead of hand-crafted heuristics
- Current MSR system:
  - Learn lexicalized transfer rules
  - No customization of parsing/generation
  - Current results: slightly better than non-customized system
- Current SYSTRAN system:
  - All components are customized

References

- Bernth, A. 1999. 'A confidence index for machine translation'. TMI-99, Chester, England.
- Jacquemin, Christian. 2001. *Spotting and Discovering Terms trough Natural Language Processing.* The MIT Press.
- Richardson, S., W. Dolan, A. Mezenes, J. Pinkham. 2001. "Achieving commercial-quality translation with example-based methods". *MT Summit VIII*, Santiago de Compostela, Spain.
- SAE J2450 Translation Quality Metric:*http://www.sae.org/technicalcommittees/j2450p1.htm*
- Rémi Zajac and Jean Senellart . 2002. 'Challenges in MT customization on closed and open text styles'. *LREC 2002 Workshop on Customizing Knowledge in NLP Applications.* Las Palmas, Spain.
- Rémi Zajac, Elke Lange and Jin Yang. 2003. 'Customizing Complex Lexical Entries for High-Quality MT'. *MT Summit IX*, New-Orleans.