Open Sources And Foreign Media Analysis

Translation at a Global Scale

Kalev Leetaru

Brief Introduction to Translation Processes

Human vs Machine Translation

- **Humans: Slowest/Highest Quality**
- **Machines: Fastest/Cheapest**

Translation Workflows

- **Fully Human**
- **Machine-Assisted Human**
- **Fully Automated**

Machine Translation

- **Expert/Rules-based. (Systran)**
- **Statistical. (LanguageWeaver)**
- **Google Translate started with Systran's rules-based translation and now offers languages such as Estonian via Statistical Translation.**

Expert/Rules-Based

- **Linguists hand program translations for every word, idiomatic expressions, grammatical rules, etc.**
- **Can take years.**
- **Highest quality for the rules they program.**
- **Fails in areas they didn't program it for (evolving expressions, etc).**

Statistical

- **Give computer Alice in Wonderland in English and in French.**
- **Determines that “bonjour” appears everytime “hello” appears.**
- **“Ventilateur” appears for “electric fan” but not “baseball fan”.**
- **Builds statistical models that tell it the probability an English word is any given French word based on the surrounding words.**

Statistical

- **“Holy Grail” of computer translation. Give it lots of text and it just learns on its own, like a human. No training required.**
- **The more text the better. Can train different translators for different time periods of text.**
- **Can handle niche languages where rules-based is more difficult.**
- **Computers just in last 5 years have enough memory to handle the models.**

Statistical

- **User feedback is critical to improving models.**
- **Google has new “Translator Toolkit” to assist: http://translate.google.com/toolkit**

Hybrid

- **Statistical is fast to train, but grammar is poor.**
- **Expert has perfect grammar, but long and expensive to train.**
- **Hybrids use statistical to translate the document and then apply rules-based to restructure the grammar to polish it.**

Translation Speed

Realtime vs Offline

- **Competing Goals: Accuracy and Speed**
- **Realtime Translation: Need it as its being said (speeches, meetings, critical announcements). Generally can't refer to dictionaries or other tools during translation task.**
- **Offline Translation: Quality is highest concern, can take more time and use external tools.**

Realtime Machines

- **Machines are excellent for realtime translation. Lesser quality, but can deliver as its being said.**
- **Mobile phone apps now offer on-the-go realtime translation: Pick a language, speak in English and it translates and speaks back in the target language.**
- **TUNGS Voice Translator for Android phones: http://www.youtube.com/watch?v=dy9l5dMigZM**

Applications: Research

Constitutions Project

- **Fill out a 1200 question survey on every written constitution in the world since 1789.**
- **English translations available for many, but rare for earlier documents, and translations may be questionable.**
- **Human quality control CRITICAL.**
- **Google Spanish translation of “never” translated into “always” in one case.**

Constitutions Project

- **We find that using machine first pass, with human correction, combines speed of machine with accuracy of human.**

Applications: Government

National Virtual Translation Center

- **Web-based portal**
- **Get back "general gist" machine translation quickly to determine if document is relevant.**
- **Can resubmit for human translation as necessary.**
- **Web portal hides the network of humans and machines.**

National Virtual Translation Center

- **Good model, lets you add/subtract translators as needed based on load.**
- **Good mix of machines for self-service translation and humans for high-accuracy work.**

BBN Technologies

- **Fully automated transcription and translation system.**
- **Allows non-linguists to have a general-gist understanding of what's being said around them (local cleric calling for protests at 10PM).**
- **1-2 second transcription delay**
- **5 second translation delay**
- **Deployed at Forward Operating Bases**

Applications: OSI

Open Source Intelligence

- **Monitor the world's news to understand what's happening where.**
- **Spies are limited resources. Far easier to monitor the news and news is very rich.**
- **Factual news (events happening) and trends (tone, cultural).**
- **Entertainment content. Cuban soap operas (new themes vs past themes).**
- **80% of actionable USSR intelligence from OSI**

Tone + Facts

- **Not just the factual translation, have to preserve tone and nuance.**
- ***Japanese radio intensifies still further its defiant, hostile tone; in contrast to its behavior during earlier periods of Pacific tension, Radio Tokyo makes no peace appeals. Comment on the United States is bitter and increased; it is broadcast not only to this country, but to Latin America and Southeastern Asia.***

Foreign Media Analysis

- **Its not just translating media, its understanding it.**
- **Cuban soap operas are one of the richest sources of info on that country right now, watching what topics are newly touched on in the shows and what topics aren't addressed anymore.**
- **Much like traditional field intelligence: get a baseline "feel" and then watch for shifts away from that baseline.**

FBIS

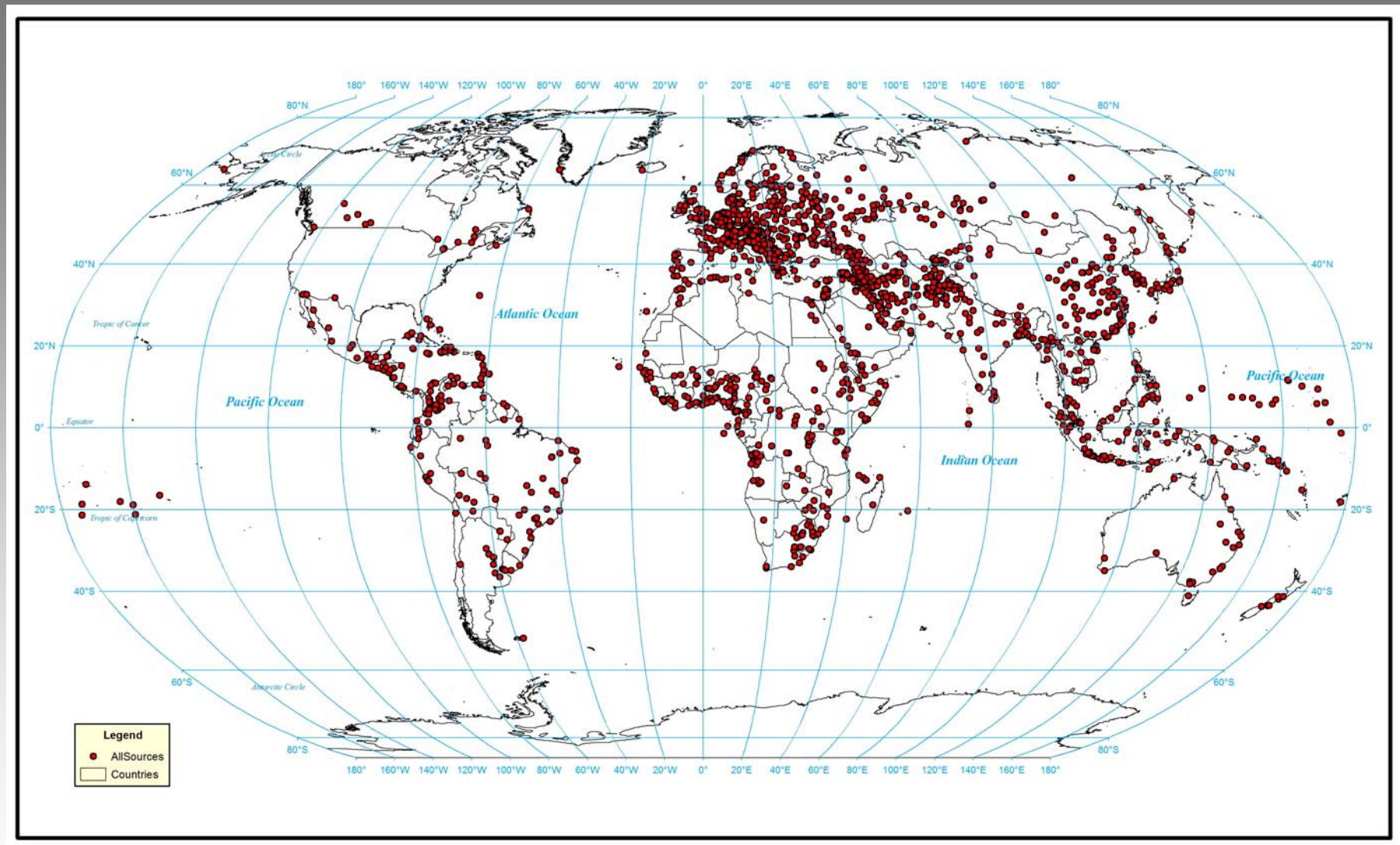
- Foreign Broadcast Information Service.
- Run by US CIA.
- Launched in 1941, continues through present.
- 1M words per day from 70+ languages.
- Broadcast (TV and Radio), Print, Internet.
- Most unclassified. Analyses classified.
- Single largest source of translated material in the world.

JPRS

- **Became component of FBIS.**
- **Focuses on technical and academic content. Will translate entire 50 page report on mining techniques from North Korea.**

SWB

- Summary of World Broadcasts
- Partner run by UK.
- Split up the world. US focuses on South America, SWB on Europe, etc.



CLINE CENTER FOR DEMOCRACY

Cline Center for Democracy at the University of Illinois

Languages

Origin Language	Report Count	% All Reports
English	2021021	46.00
Russian	371106	8.45
Arabic	271326	6.18
Spanish	197451	4.49
French	138046	3.14
Serbo-Croatian	135805	3.09
Chinese	124014	2.82
Persian	80720	1.84
German	76688	1.75
Portuguese	66003	1.50
Turkish	65951	1.50
Hebrew	51670	1.18
Japanese	50509	1.15
Korean	47113	1.07
Albanian	40898	0.93
Italian	39060	0.89
Urdu	31705	0.72
Ukrainian	31608	0.72
Indonesian	29359	0.67
Greek	28564	0.65
Polish	28372	0.65
Hungarian	26392	0.60
Slovak	22980	0.52
Bulgarian	22920	0.52

English

- **Lots of news sources offer pretranslated content. Often for expats or foreign investors.**
- **Convenient, but huge risks. Always have to understand your source.**
- **ITAR-TASS carefully selects the stories it translates and often adds/leaves out portions or changes tone.**

Coverage Networks

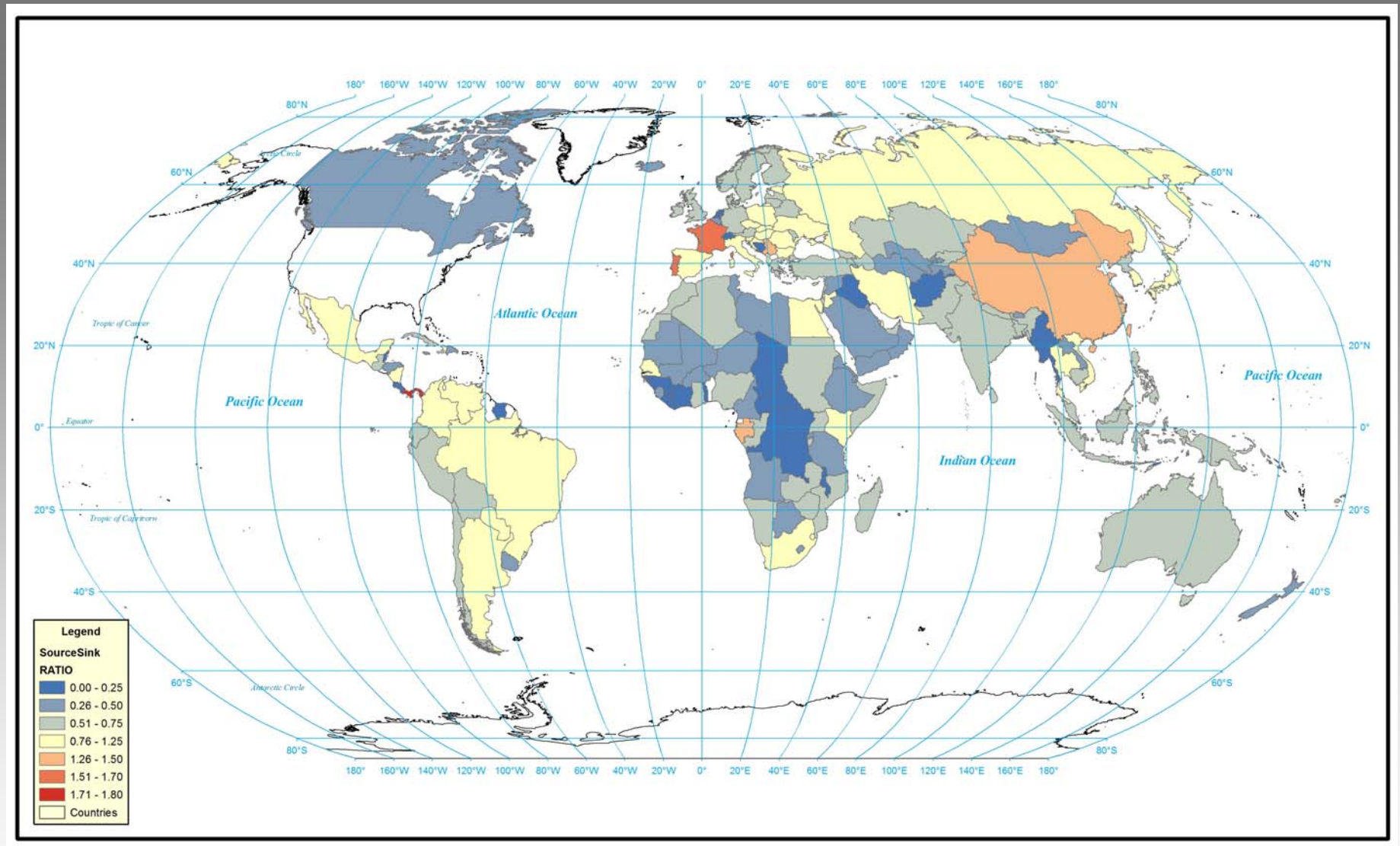
- **Often you can't get penetration into local media because it is non-existent, difficult/impossible to access (local low-power rural transmitter), etc.**
- **Use neighboring countries, or those with strong cultural or economic ties.**

Coverage Networks

- **News: Russian media covers ex-Soviet states in far greater detail than Brazil does. Use Russia for coverage of hard-to-reach states.**
- **Documents: French libraries have rich archives of African legal journals. Most African libraries have not fully digitized their card catalogs, but French libraries have. Use as proxy.**

Coverage Networks

Source	Report Count	% All Reports
Beijing XINHUA	194316	4.42
Moscow ITAR-TASS	155925	3.55
Tokyo KYODO	123404	2.81
Seoul YONHAP	92722	2.11
Tehran IRNA	57857	1.32
Paris AFP	56286	1.28
Hong Kong AFP	44390	1.01
Prague CTK	39201	0.89
Ankara Anatolia	31436	0.72
P'yongyang KCNA	29824	0.68
Moscow INTERFAX	29141	0.66
Belgrade BETA	28717	0.65
Belgrade TANJUG	28381	0.65
Cairo MENA	26764	0.61
Pyongyang KCNA	26230	0.60
Zagreb HINA	23013	0.52
Taipei Central News Agency WWW-Text	22983	0.52
Moscow RIA	22071	0.50
Tokyo Jiji Press	21508	0.49
Moscow Nezavisimaya Gazeta	20371	0.46
Moscow Agentstvo Voyennykh Novostey WWW-Text	19931	0.45
Jerusalem Qol Yisra'el	19896	0.45
Madrid EFE	18973	0.43
Warsaw PAP	17903	0.41


Atlantic Ocean
Pacific Ocean
Pacific Ocean
Indian Ocean
Tropic of Cancer
Equator
Tropic of Capricorn
Antarctic Circle
Legend
SourceSink
RATIO
0.00 - 0.25
0.26 - 0.50
0.51 - 0.75
0.76 - 1.25
1.26 - 1.50
1.51 - 1.70
1.71 - 1.80
Countries

Workflow

- **Entirely human-based translation.**
- **Human and machine-based categorization.**
- **Translation products are sent out on wire, downstream consumers can request clarification on particular terms (was this “no” or “never”?), and translations revised/enhanced and rereleased.**

Iterative Workflow

- In mission critical environments, translation becomes an iterative process.
- You provide translation to users.
- Users may need clarifications or special accuracy on sections.
- Retranslate those parts and resubmit.
- Broadcast content in particular requires revisions.

Thank you!

- **Kalev Leetaru**
  - leetaru@illinois.edu


CLINE CENTER
FOR DEMOCRACY