OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

MATERIAL

Designing an MT Program for IARPA

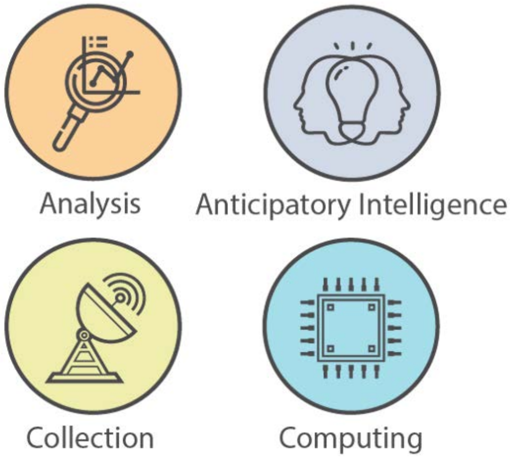
Dr. Carl Rubino
IARPA

Intelligence Advanced Research Projects Activity

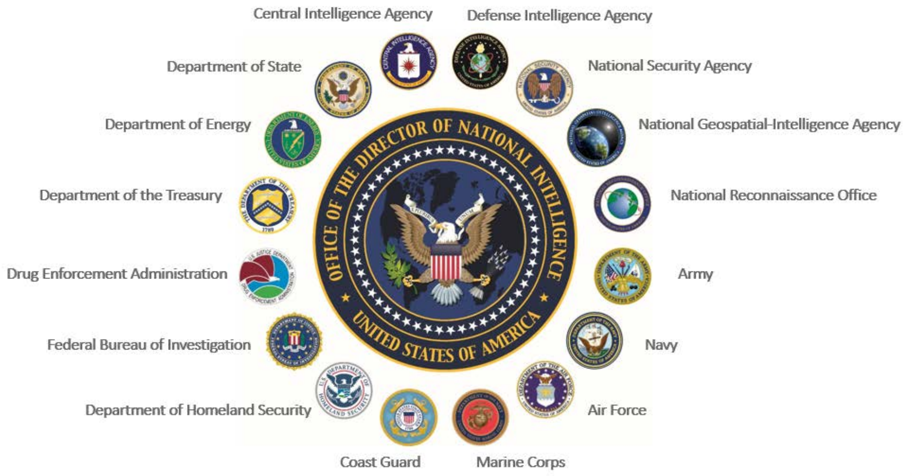
IARPA envisions and leads *high-risk, high-payoff research* that delivers innovative technology *for future overwhelming intelligence advantage*

- Our problems are **complex** and **multidisciplinary**
- We emphasize **technical excellence** & **technical truth**

4 Core Research Thrusts



The United States Intelligence Community


Central Intelligence Agency
Defense Intelligence Agency
Department of State
National Security Agency
Department of Energy
National Geospatial-Intelligence Agency
Department of the Treasury
National Reconnaissance Office
Drug Enforcement Administration
Army
Federal Bureau of Investigation
Navy
Department of Homeland Security
Air Force
Coast Guard
Marine Corps
OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE
UNITED STATES OF AMERICA

IARPA does everything "from AI to Zika" and is a world scientific leader

Although best known for quantum computing, superconducting computing and forecasting tournaments – IARPA's research portfolio is diverse, including math, physics, chemistry, biology, neuroscience, linguistics, political science, cognitive psychology and more.

- **70% of completed research transitions** to U.S. Government partners
- **2,000+ journal articles** published through FY2016
- Physicist David Wineland won the **Nobel Prize in Physics** for quantum computing research funded by IARPA
- World's leading funder of quantum computing academic research, and quantum research cited as Science Magazine's "Breakthrough of the Year"
- White House BRAIN Initiative, National Strategic Computing Initiative
- Dr. Craig Gentry named a **MacArthur Fellow**

Program Impetus

ახალი ცოცხი კარგად ჰგვის, ძველი _ მოატანს ქვიშასა.

ለሆዳም ሰው ማብላት ውቅያኖስን ለመደልደል መቃጣት

No me olvides.

东南西北 方方 福星高照

ගිය ලූලා මහා එකා ලූ

Ang sakit ng kalingkingan, sakit ng buong katawan

શ્રદ્ધા કપૂરના પાત્રની વિગતો જાહેર થઇ

Kuelekeza si kufuma, na kuchumbia si kuoa

नजाने गाउँको बाटै नसोध्नु ।

미국 투자이민의 생각이 바뀐다

Утопáющий за солóминку хватáется.

ęŇĹŢĚŽdI·ĂđĐ

MATERIAL Goal

- Revolutionize multilingual triage by enabling rapid development of language-independent methods (when possible) to field systems capable of fulfilling domain-specific cross-language information retrieval tasks over both text and speech data, with:
  - Limited bitext and transcribed speech training data
  - English domain-specific queries as input
  - English summaries of retrieved results as output
  - Methods for domain adaptation and portability to new languages
  - Assessment of the technology via a resonating end-to-end use case

The MATERIAL System

- An “English-in, English-out” information retrieval system that, given a domain-sensitive English query, will retrieve relevant data from a large multilingual repository and display the retrieved information in English as summaries that reflect the document relevance:

English query in

(specifying
WORDS = “Robot”;
DOMAIN= “Military”)

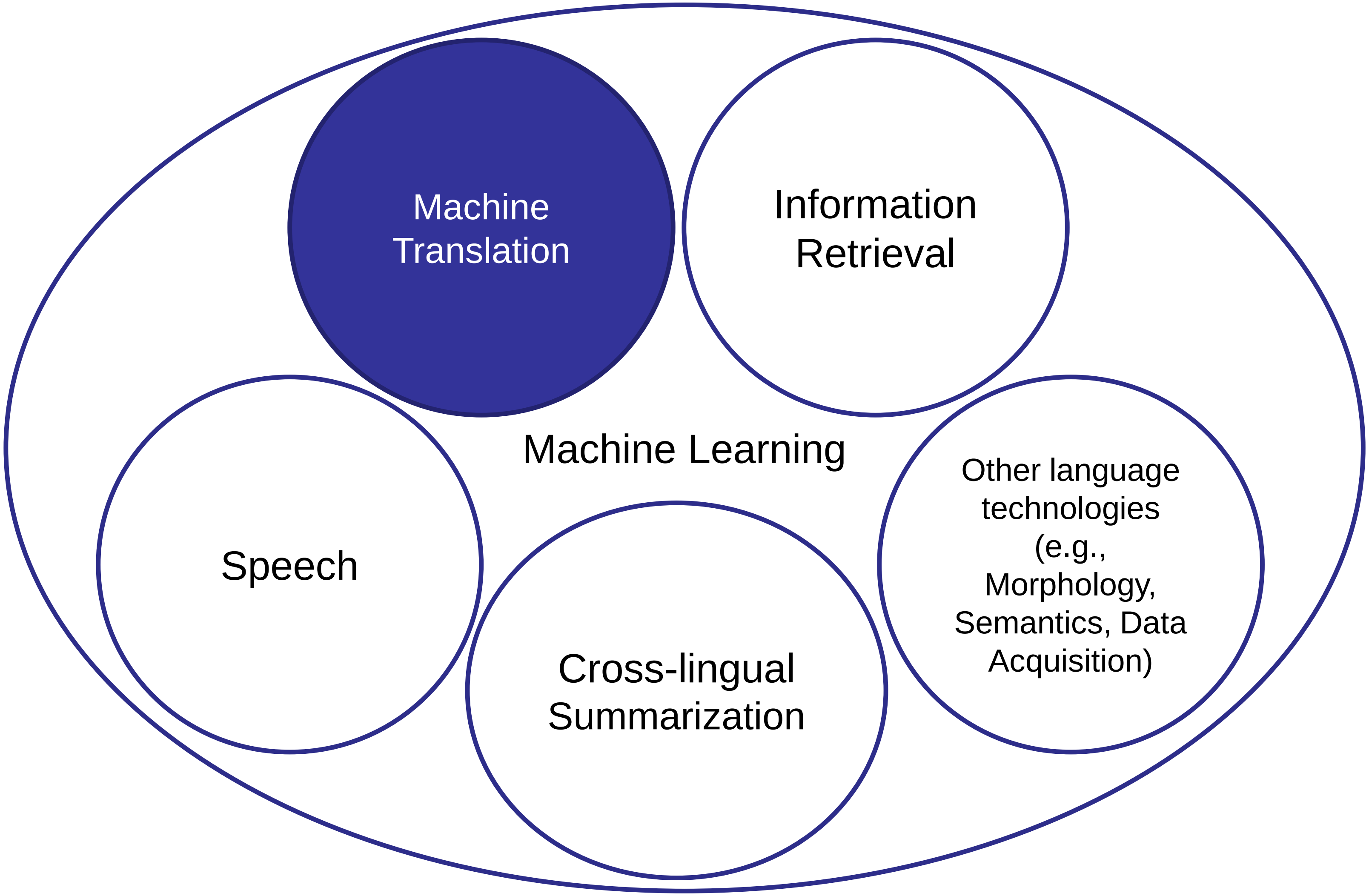
Robots stormed IARPA
로봇은 지능 고급 연구 프로젝트 기관을 습격
እንዲፈትሽ ያስገድዳሉ። መንግሥት ውስጡን
اقتحم الروبوت الاستخبارات
وكالة المتقدم مشروع بحث
Робот штурмовали разведки Агентства перспективных исследований проекта
εμπιστοσύν ης στην κυβέρνηση είναι
Y, por ma que enten distraerlo, este humanoide diseñado por una empresa robótica de Google
หุ่นยนต์บุกสำนักงานข่าวกรองโครงการวิจัยขั้นสูง

English Summary Out

俄罗斯国防工业部门最新推出的“铀-6”扫雷机器人极

Russian defense industry uranium 6 minesweeper robot

MATERIAL Technical Areas



Machine Translation

Information Retrieval

Machine Learning

Speech

Cross-lingual Summarization

Other language technologies (e.g., Morphology, Semantics, Data Acquisition)

Why Low Resource Languages?

ಅಂಕಾರಾ: ಟರ್ಕಿ ರಾಜಧಾನಿ ಅಂಕಾರದಲ್ಲಿ *ಮಿಲಿಟರಿ* ಪಡೆಯನ್ನು ಗುರಿಯಾಗಿರಿಸಿ ಬುಧವಾರ ಸಂಭವಿಸಿದ ಕಾರು ಬಾಂಬ್ ದಾಳಿಯಲ್ಲಿ 28 ಮಂದಿ ಸಾವಿಗೀಡಾಗಿದ್ದು, 61 ಮಂದಿಗೆ ಗಾಯಗಳಾಗಿವೆ. *ಮಿಲಿಟರಿ* ವಾಹನಗಳು ಹಾದು ಹೋಗುತ್ತಿರುವ

Military convoy

Car Bomb

- Fundamental Challenges
  - Lack of bitexts or commercial products to exploit
  - Incomplete grammatical descriptions
  - Scant lexical resources
  - May have radically divergent typologies and complex morphologies
  - May have encoding anomalies
  - Segmentation issues
  - Lack of orthographic standardization; Multiple scripts
  - Informal genres of text will capture orthographic deviation (typos vs. truncations)

Query Format

- **Domain-specific (e.g., Government, Lifestyle)**
- **Address domain-restricted information need**

“polio vaccine”
Domain: Government and Politics

Subject Domain

✓ ✓ …In response, the Armenian **Ministry of Health** urged all Syrian Armenians under age 15 to get the **polio vaccination**…

✗ ✗ …Severe adverse reactions to this **vaccine** are rare.…

✓ ✗ …The oral **vaccine** was made by weakening the three strains of **poliovirus** that caused disease by growing them in monkey kidney cells…

Non-traditional Query Constructs

Semantic Expansion:
“environmental protection”+
“planting okra”+
Surigao, irrigation+

Constraint Disambiguation:
Fly [hypernym: insect]
Light [event frame:weight]
“Buddha’s hand” [synonym: Citrus medica sarcodactylis]

Categorical Membership:
EXAMPLE_OF (mammal)
EXAMPLE_OF (dairy product)

Morphological Constraint:
<helicopters>

+ domain!

Key Technical Challenges

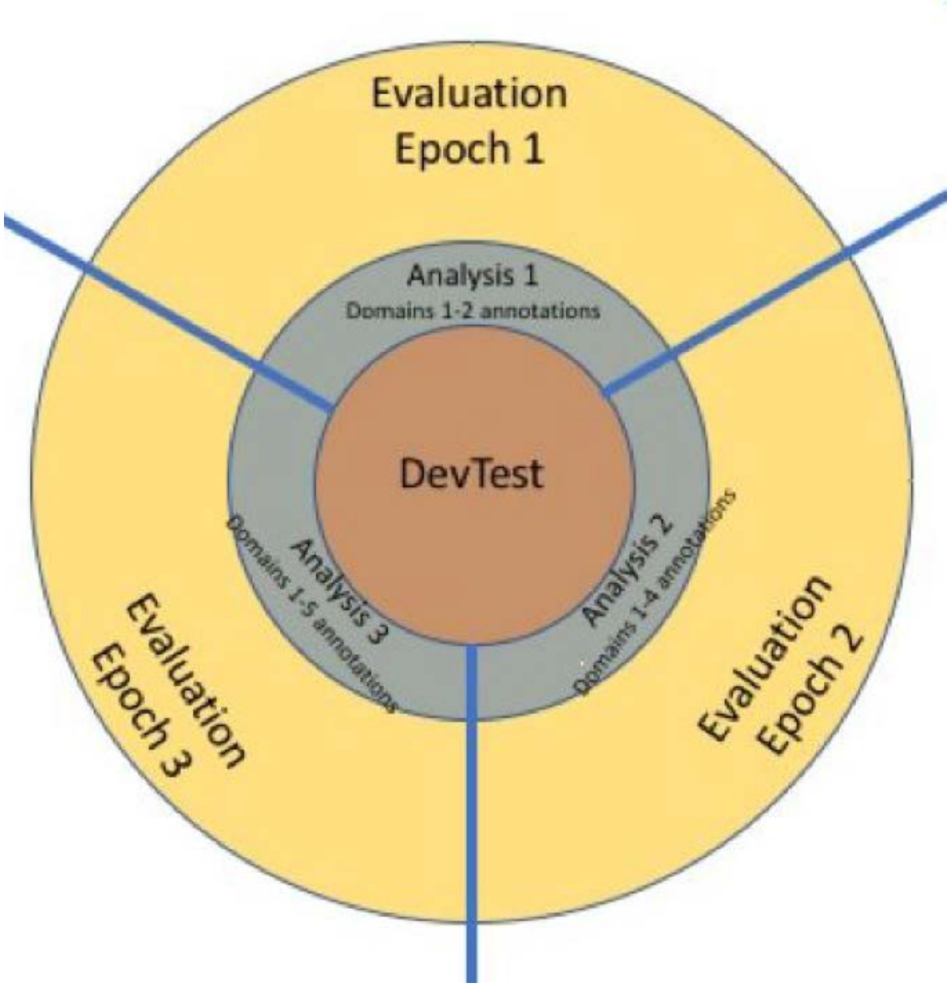
- Techniques appropriate for a wide variety of languages
- Performance on formal and informal text and speech in a wide variety of genres that do not match the training conditions
- Development of new methods for domain adaptation without monolingual or parallel training data in that domain
- Limited time to develop a fully automatic E2E system to process a new language
- Inter-language domain mismatches reflecting a cultural component
- Use of web resources to complement limited training data

Performers must develop methods that are not data intensive.

MATERIAL Training Data

- Each language will be provided at kick-off to performers in a “pack” from the IARPA T&E Team that will contain training data for MT and ASR as well as relevant language information
- Speech data will include:
  - Roughly 50 hours of conversational speech transcriptions in a *normalized orthography* in two demographically balanced partitions (40 hours training; 10 hours dev test)
  - Phonetic Lexica
  - Not all genres and domains will be present in the training data
- MT Training data will include:
  - 800K word bitexts in each program language, sentence-aligned. Up to five sentences may be included with a denoted grouping. Translations are purely human and new.
  - Not all genres and domains will be present in the training data
- Language information will include:
  - Description of the language (e.g., dialect regions, phoneme set definitions)
  - Basic information on dialects, spelling and encoding

MATERIAL Data Partitioning



Query Releases:

QR1: 2 Domains – Open
QR2: 4 Domains – Closed
QR3: 5 Domains – Closed

Build Packs for Training:

ASR Training & Phonetic Lexica
MT Bitexts

Domains and Genres Used for Development and Evaluation

Program data will include formal and informal varieties of text and speech, including genres that are not present in the MT or ASR training data.

Mode	% Collect	Genre
Text	~ 75	News
		Topical
		Social Media
Speech	~ 25	Broadcast News
		Topical Broadcasts
		Conversation

Domains (Broad Subject Fields) will vary from language to language based on corpus characteristics, e.g. Government and Politics, Health, Military, etc.

Test Structure Rationale

- T&E regimen designed to drive R&D towards the program goal, viz:

  Language independent methods, tools, and technologies to provide **rapid-deployment** of **domain-adapted** MT for **low-resource** languages effectively integrated in a usable CLIR system

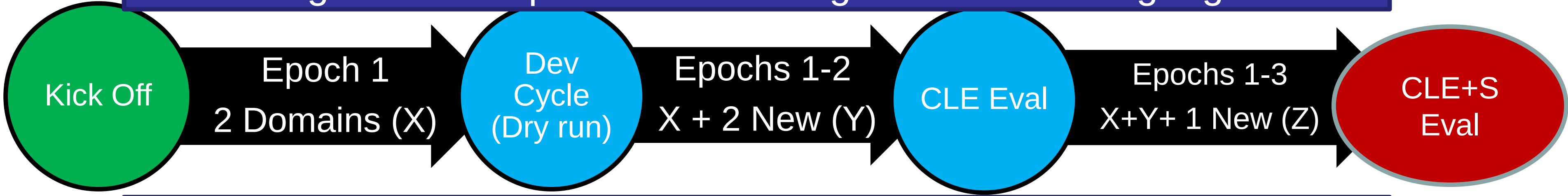
- So:
  - Multiple languages with varying characteristics
  - Only small amounts of IARPA-furnished bitexts for training
  - Domain contextualized queries
  - Decreasing lead-time for development & surprise language evaluation

	Base	Option 1	Option 2
Length (months)	18	16	12
# Practice Languages	2	2	3
Surprise Language Period (months)	6	4	1.5

Program at a Glance

Training Data at Each Kickoff Period per language: 800K Words Bitexts; 50 Hours Transcribed Audio

Stage 1: Development and Testing on Practice Languages



Kick Off

Epoch 1
2 Domains (X)

Dev Cycle (Dry run)

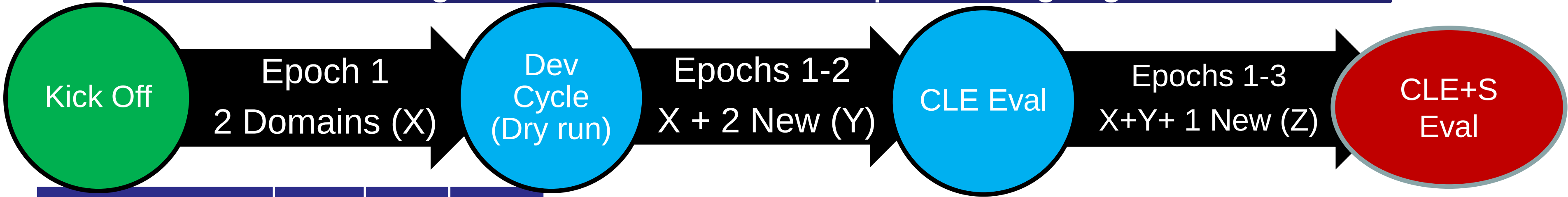
Epochs 1-2
X + 2 New (Y)

CLE Eval

Epochs 1-3
X+Y+ 1 New (Z)

CLE+S Eval

Stage 2: Evaluation on 1 Surprise Language



Kick Off

Epoch 1
2 Domains (X)

Dev Cycle (Dry run)

Epochs 1-2
X + 2 New (Y)

CLE Eval

Epochs 1-3
X+Y+ 1 New (Z)

CLE+S Eval

	Base	Opt 1	Opt 2
# Dev Languages	2	2	3
Phase Duration	18	16	13
Practice CLE	7	8	8
Practice CLE+S	10	10	9
Surprise CLE	13	12	10.5
Surprise CLE+S	16	14	11.5
Surprise duration	6	4	1.5

Staging in of Queries:

Kickoff	Test 1	Test 2
50% X queries	75% X & 50% Y queries	100% X, Y & Z queries

10% responsive data translations provided at each stage for analysis that cannot be used for further training.

CLIR Detection Metric: AQWV
Actual Query Weighted Value

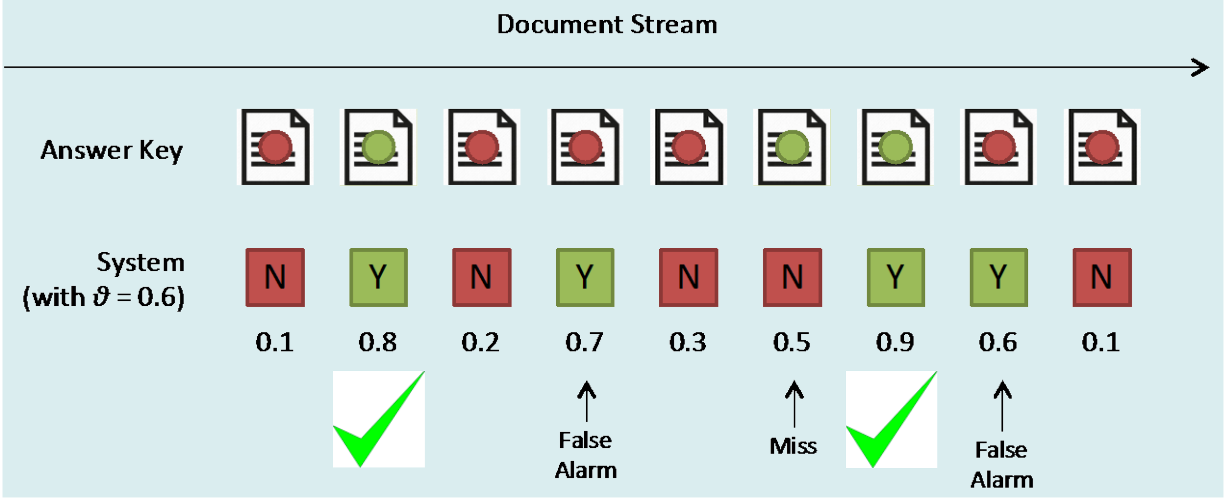
- All queries are treated equally (regardless of whether they generate single or multiple hits).
- Must be able to calibrate the metric against a baseline CLIR system that has two inputs: GOTS MT and human translation.
- Metrics will be reported to performers as an average over the set of queries (not individually for each query).
- Calculated as a representation of error rate taking into account probability of hits, false alarms, and the total number of responsive documents. These parameters will be set by T&E once the data are collected and evaluated.

Actual Query Weighted Value (AQWV)

Developers choose Θ, the detection threshold for their “Actual Decisions”, to optimize query-weighted value

- V is the *a priori* value (benefit) of a correct response
- C is the *a priori* cost of an incorrect response
- P_{rel} is the prior probability that a document is relevant to a query, e.g. 10^{-3}

$$Value_Q(\theta) = 1 - \underset{Q}{\text{average}}\left\{p_{miss}(Q,\theta) + \frac{C}{V}\left(p_{rel}^{-1} - 1\right)\cdot p_{fa}(Q,\theta)\right\}$$



Developers will tune their systems to the threshold that maximizes the AQWV. Note that 1 is a perfect score; that is, error rate is zero.

Evaluating Summarization

CLIR Contingency Matrix

		Y	N
Key	Y	X_1 True Positive	X_2 False Negative
	N	X_3 False Positive	X_4 True Negative

$$QWV = 1 - \frac{X_2}{(X_1 + X_2)} - \beta \frac{X_3}{(X_3 + X_4)}$$

Crowd Summary Judgments

		Y	N
CLIR/ Key	Y/Y (=X_1)	A	B
	Y/N (=X_3)	C	D

A "perfect" summarization capability would hold B at zero and reduce C to zero

CROWD-SOURCED JUDGMENTS:

Y/Y: Retrieved docs that are relevant
Y/N: Retrieved docs that are not relevant

A : # Relevant docs judged relevant
B : # Relevant docs judged non-relevant
C : # Non-relevant docs judged relevant
D : # Non-relevant docs judged non-relevant

Evaluating Summarization (cont.)

CLIR Contingency Matrix

		Y	N
Key	Y	X_1 True Positive	X_2 False Negative
	N	X_3 False Positive	X_4 True Negative

$$QWV = 1 - \frac{X_2}{(X_1 + X_2)} - \beta \frac{X_3}{(X_3 + X_4)}$$

A “perfect” summarization capability would hold B at zero and reduce C to zero

Crowd Summary Judgments

		Y	N
CLIR/ Key	Y/Y (=X_1)	A	B
	Y/N (=X_3)	C	D

End to End Contingency Matrix

		Y	N
Key	Y	A	X_2+B
	N	C	X_4+D

Summarization can reduce the false alarm rate ($C \leq X_3$) but cannot reduce the number of missed detections ($B \geq 0$)

$$QWV = 1 - \frac{X_2 + B}{(A + X_2 + B)} - \beta \frac{C}{(C + X_4 + D)}$$

T&E (Test and Evaluation)

- The Data

Data and/or annotation were supplied from four primary sources: Appen Butler Hill, Inc. (Appen), The Center for Applied Machine Translation at NSA, The National Virtual Translation Center at FBI, and Air Force Research Laboratories (AFRL).

- The T&E Team

Test & Evaluation

MIT Lincoln Laboratory

NIST National Institute of Standards and Technology

University of Maryland CASL Center for Advanced Study of Language

Tarragon Consulting Corporation

T&E Team Roles

UNIVERSITY OF MARYLAND
CASL
CENTER FOR ADVANCED
STUDY OF LANGUAGE

- The Center for the Advanced Study of Language (CASL) provides guidance on the languages, data quality, annotation and linguistic aspects of evaluation.

  POCs: Anne David and Aric Bills

NIST
National Institute of
Standards and Technology

- The National Institute of Standards and Technology (NIST) is the evaluation lead for the T&E team. They design and administer the CLIR test sets via their test server.

  POCs: Greg Sanders and Audrey Tong

T&E Team Roles

MIT Lincoln Laboratory

- MIT Lincoln Laboratory (LL) supports the program with data collection, data annotation, partitioning and vetting. They have built a baseline system to better understand program challenges and appropriately set expectations.

  POCs: Nick Malyska and Jennifer Melot

Tarragon CONSULTING CORPORATION

- Tarragon Consulting provides ontological support for the program, crucial for our data annotation contract effort. They also lead the efforts for summary evaluation.

  POC: Richard Tong

First Two Practice Languages

- From a Broad language portfolio:
  - Different Language Families
  - Mixed language typology (i.e., with different phonotactic, morphological, syntactic characteristics)

Tagalog

Manila
Philippines
Malaysia

Swahili

Somalia
Uganda
Kenya
Rwanda
Democratic Republic of the Congo
Burundi
Tanzania
Zambia
Malawi
Mozambique
Madagascar

Swahili (1A)

- SVO Word Order
- Agglutinative Morphology with prefixes and suffixes
- Verbs and adjectives agree with noun class of subjects and objects
- 18 noun classes
- Latin script, 5 vowels, 33 consonants (4 pre-nasalized, 4 borrowed)
- Variability (dialects)

Mama	anamlisha	mtoto	uji.
mama	a-na-m-lish-a	m-toto	u-ji
mother	3SG-PRES-CL1.OBJ-feed-FV	CL1-child	CL14-porridge

‘Mother is feeding the child porridge.’

Mti	mrefu	mwembamba	ulianguka.
m-ti	m-refu	m-embamba	u-li-anguk-a
CL3-tree	CL3-tall	CL3-thin	CL3-PST-fall-FV

‘The tall thin tree fell.’

Tagalog (1B)

- Predicate-initial word order, Agent usually precedes patient
- Agglutinative Morphology with prefixes, suffixes, infixes and reduplication
- Verbs take derivational affixes to denote semantic relation of NOM
- Latin script. Contrastive lexical stress. Heavy borrowing – diglossia.

- Agglutinative morphology with reduplication:
  - Prefix: mag-Tagalog = ‘to speak Tagalog’
  - Suffix: Tagalug-in = ‘say in Tagalog’
  - Infix: T[in]agalog = ‘translated into Tagalog’
  - Circumfix: ka-Tagalug-an = ‘Tagalog area’
  - Reduplication: nag-***ta***-Tagalog = “speaking Tagalog”
  - And combinations thereof
- Predicate initial language, Agent usually precedes patient

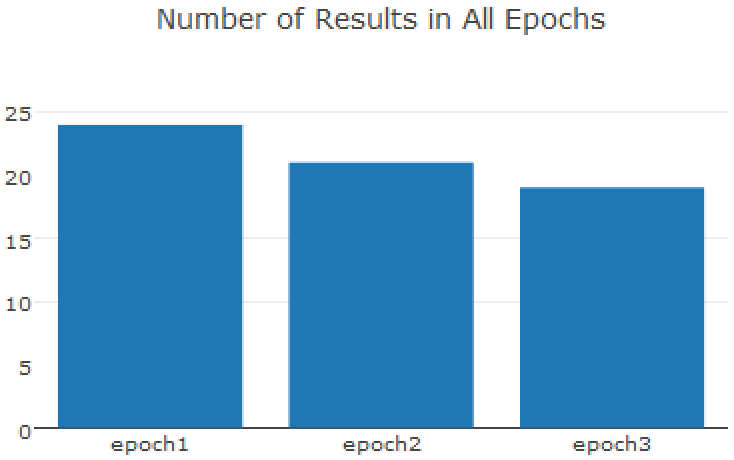
–	Inubus	ko	yung	pansit	niya.
–	finished	1s	SPEC	noodles	3s.GEN

  - ‘I finished his/her noodles’
- Traditionally a three vowel system written with five vowels
  - Word final vowel lowering: U > O, I > E / _#

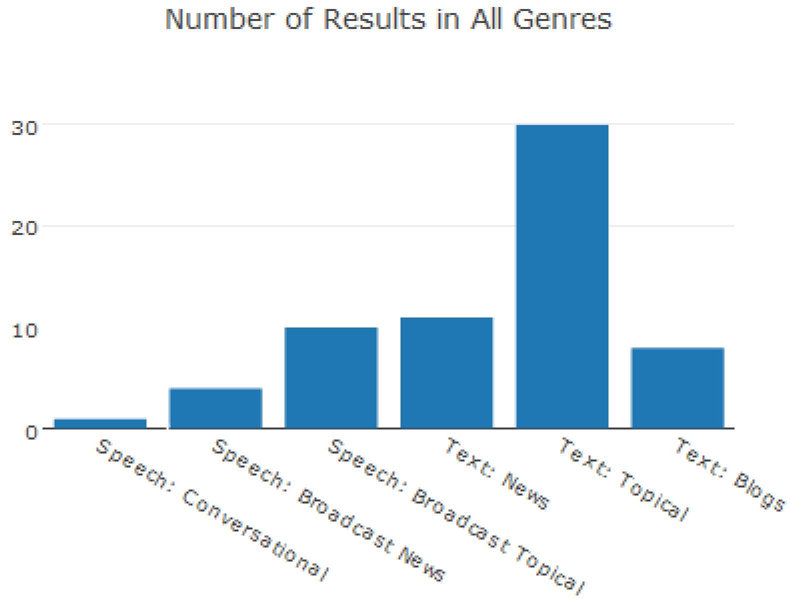
Kababaihan (from babae)
Tagalugin (from Tagalog)

Query Development

Query 1B-2034
Domain GOV
Named Entity

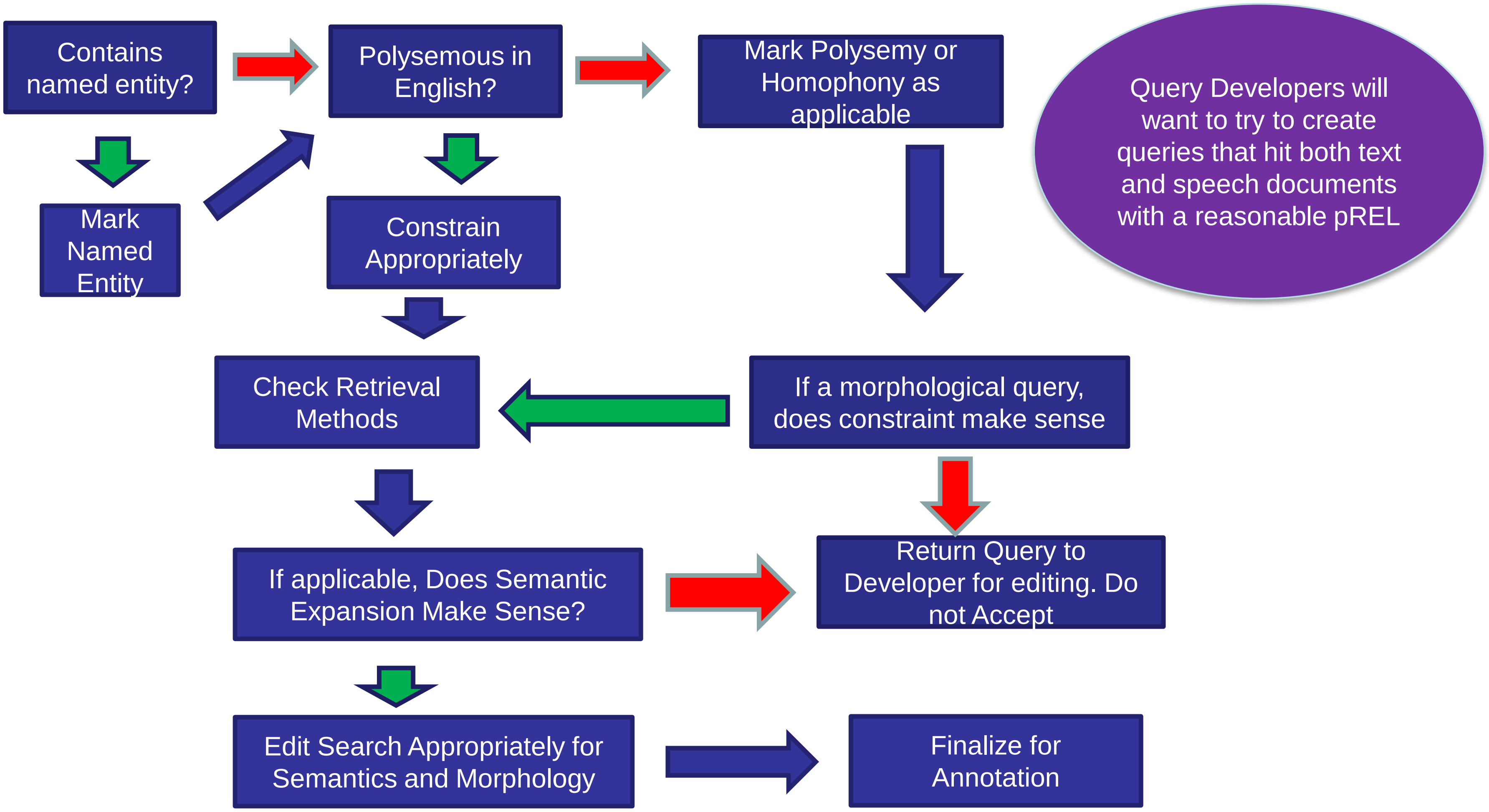


Number of Results in All Epochs



Number of Results in All Genres

Vetting Query Development



Contains named entity?

Mark Named Entity

Polysemous in English?

Mark Polysemy or Homophony as applicable

Query Developers will want to try to create queries that hit both text and speech documents with a reasonable pREL

Constrain Appropriately

Check Retrieval Methods

If a morphological query, does constraint make sense

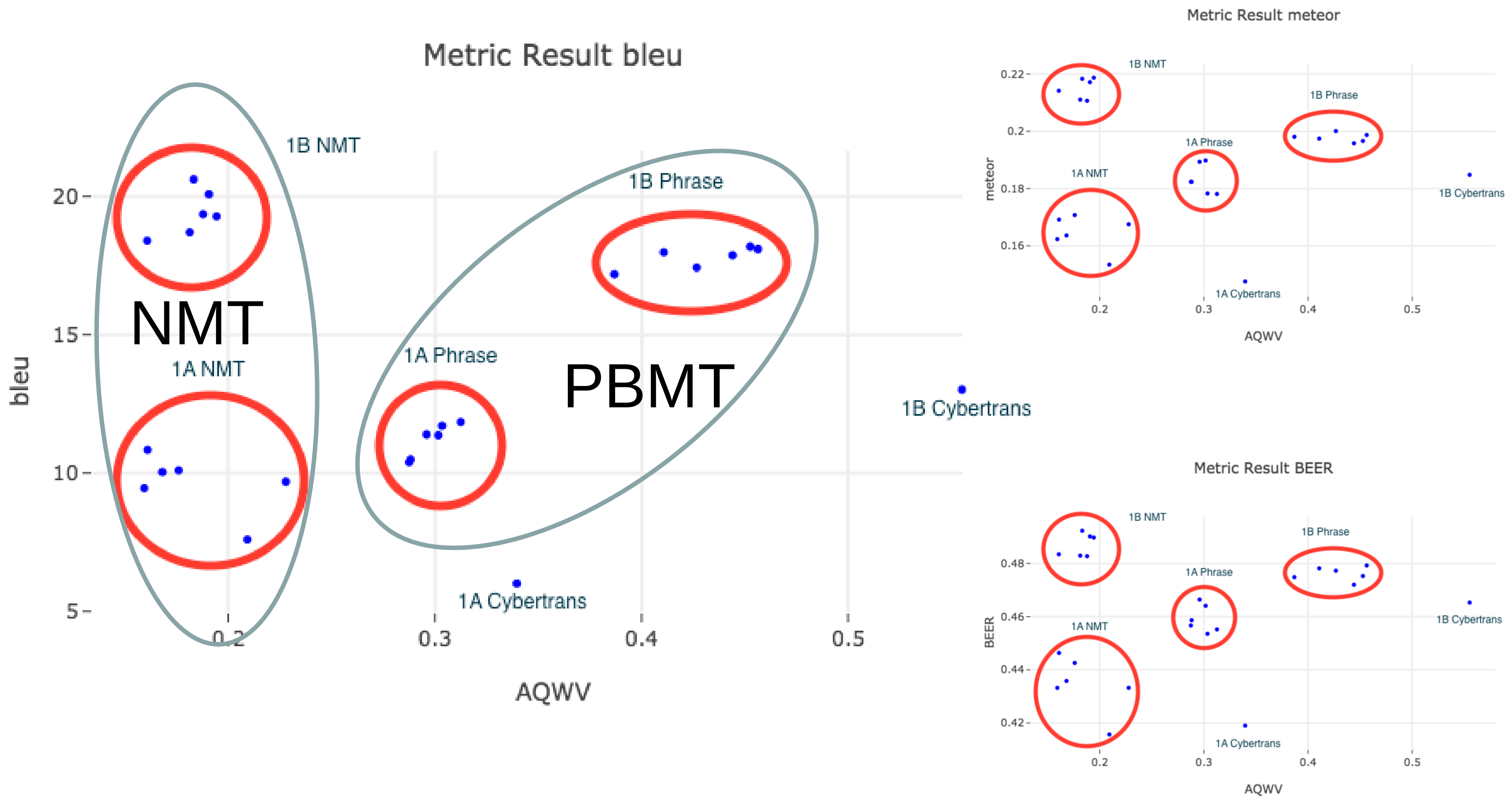
If applicable, Does Semantic Expansion Make Sense?

Return Query to Developer for editing. Do not Accept

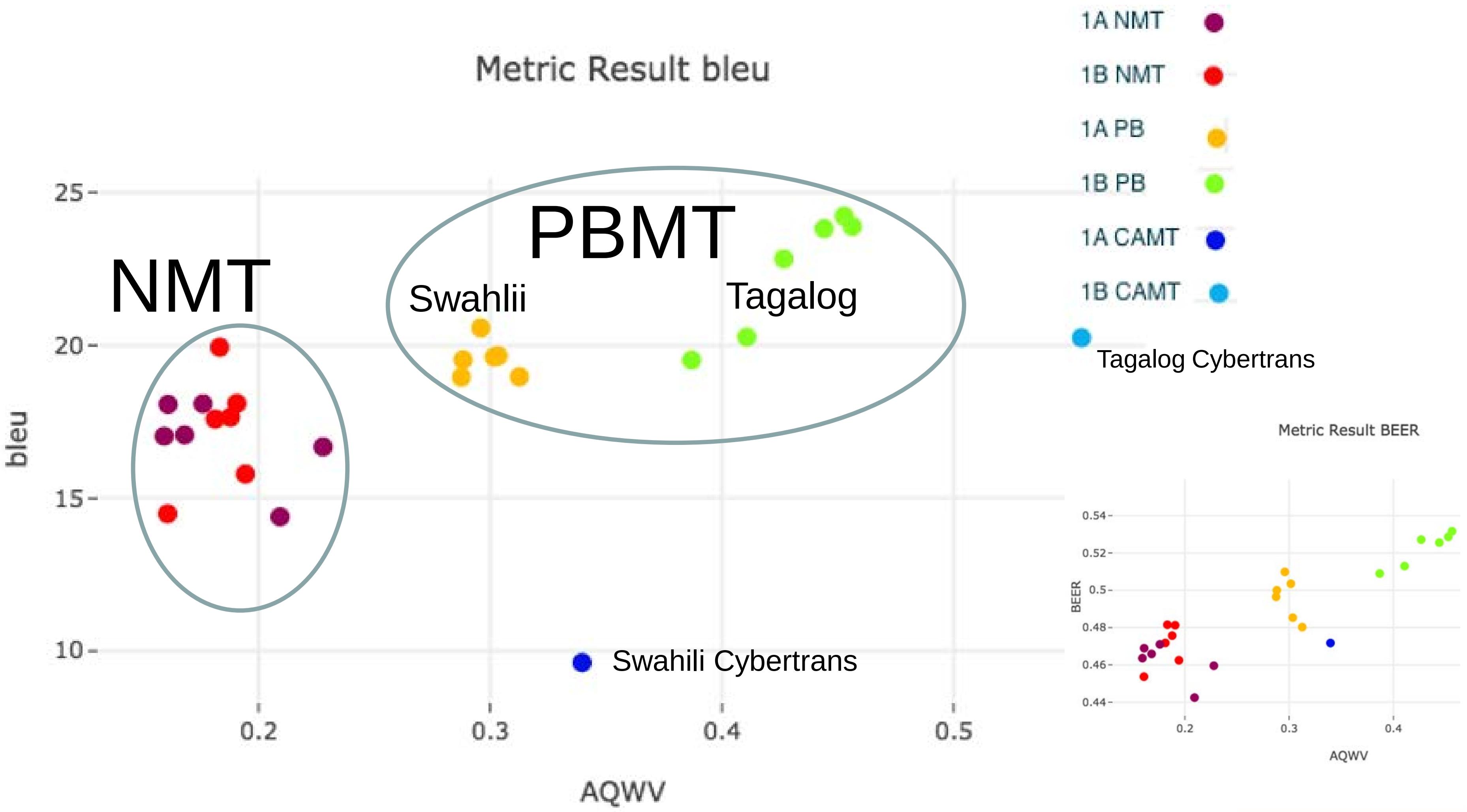
Edit Search Appropriately for Semantics and Morphology

Finalize for Annotation

Baseline Metric Correlation, Eval1



Baseline Metric Correlation, Analysis1



Four MATERIAL Performing Teams

JOHNS HOPKINS UNIVERSITY

Raytheon

Raytheon BBN Technologies

BRNO UNIVERSITY OF TECHNOLOGY

THE UNIVERSITY OF EDINBURGH

UNIVERSITY OF MASSACHUSETTS AMHERST 1863

UNIVERSITY OF WASHINGTON 1861

COLUMBIA UNIVERSITY

UNIVERSITY OF CAMBRIDGE

THE UNIVERSITY OF EDINBURGH

UNIVERSITY OF MARYLAND 1856

YALE

USC UNIVERSITY OF SOUTHERN CALIFORNIA

ISI Information Sciences Institute USC School of Engineering

idiap RESEARCH INSTITUTE

Raytheon BBN Technologies

MIT Massachusetts Institute of Technology

NORTHEASTERN UNIVERSITY BOSTON, MASSACHUSETTS

RENSSELAER POLYTECHNIC INSTITUTE 1824

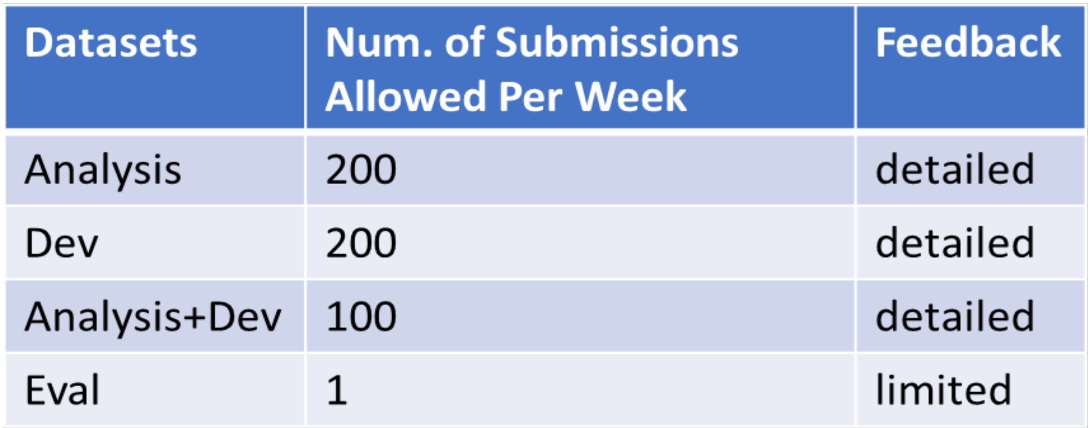
UNIVERSITY OF MASSACHUSETTS AMHERST 1863

SIGILLUM UNIVERSITATIS DOMINAE NOSTRAE A LACU

… and some initial results and discoveries

Team Scoring

- Web-based scoring server hosted by NIST
- Allows performer teams to submit their system outputs and get score feedback on various datasets



Datasets	Num. of Submissions Allowed Per Week	Feedback
Analysis	200	detailed
Dev	200	detailed
Analysis+Dev	100	detailed
Eval	1	limited

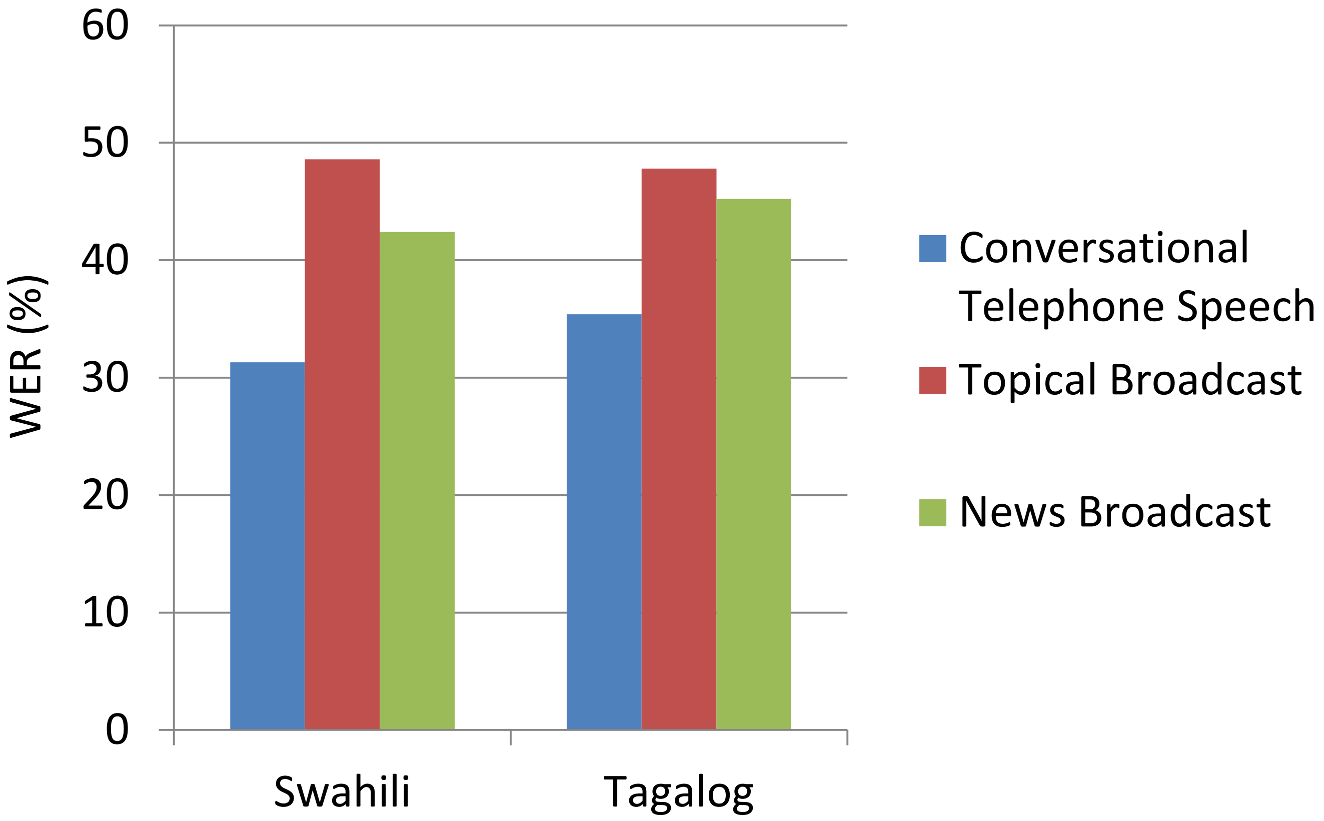
Detailed results for CLIR task include AQWV breakdown by:

- Document types (text, speech)
- Domains
- Query types (lexical, conceptual, hybrid)
- Query characteristics
    - EXAMPLE_OF
    - Semantic constraint (hypernym, synonym, event frame)
    - Morphological constraint
    - Conjunction
    - Constituent phrase
    - Multi-word phrase
    - Contains capital letter

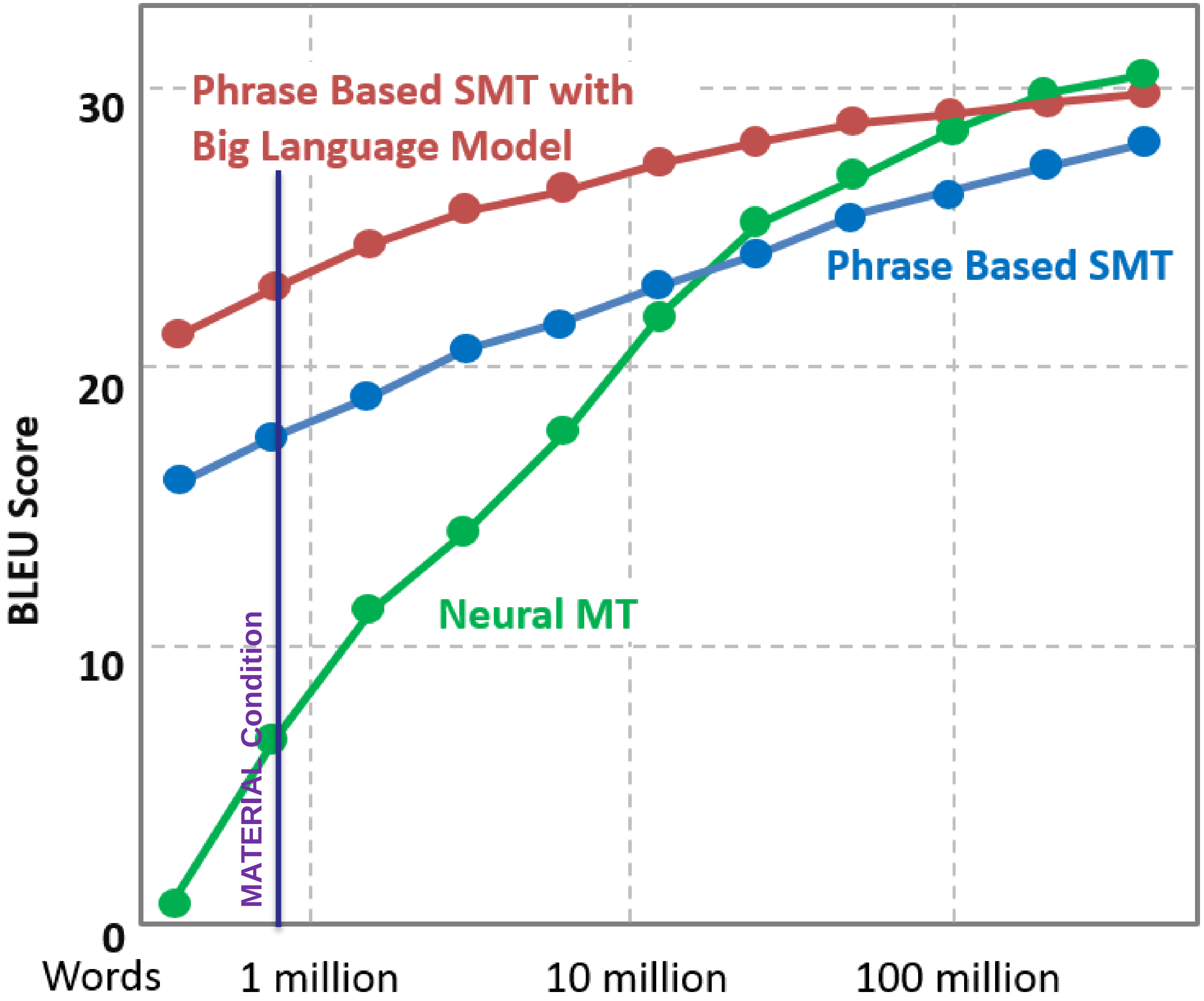
Mismatch Degradation from Analysis Docs

- Conversational vs. Broadcast Performance

Word Error Rate Metric – Lower is Better



Data Needs: SMT vs. NMT

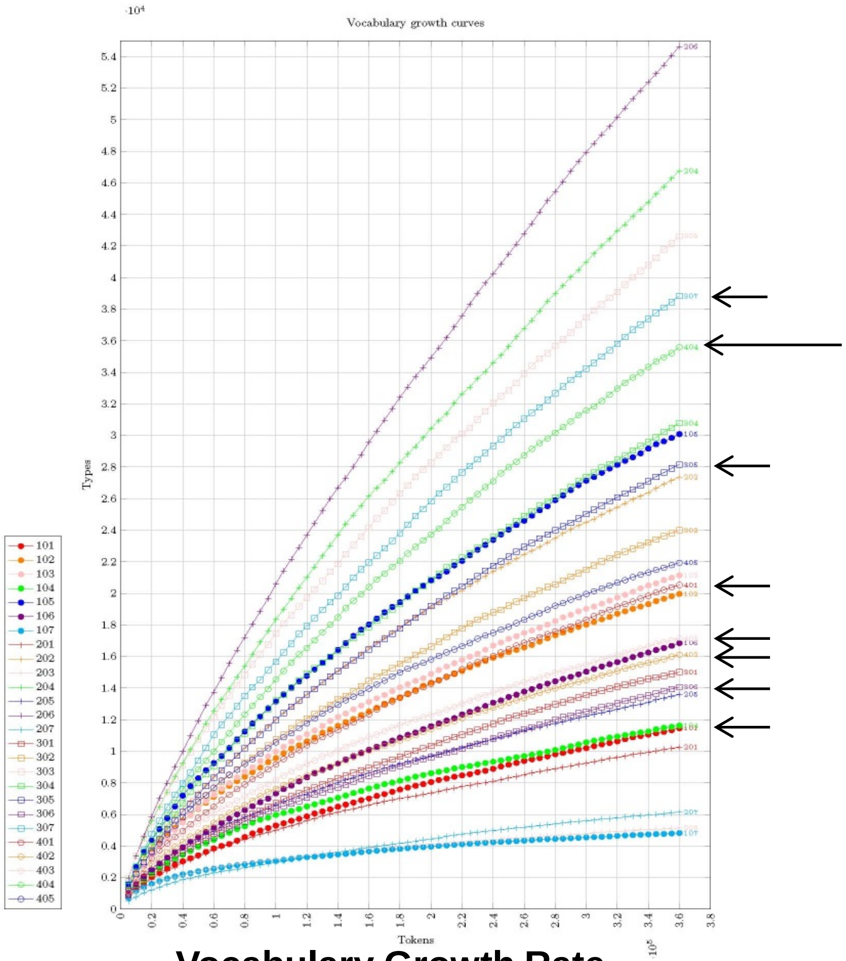


Corpus Size

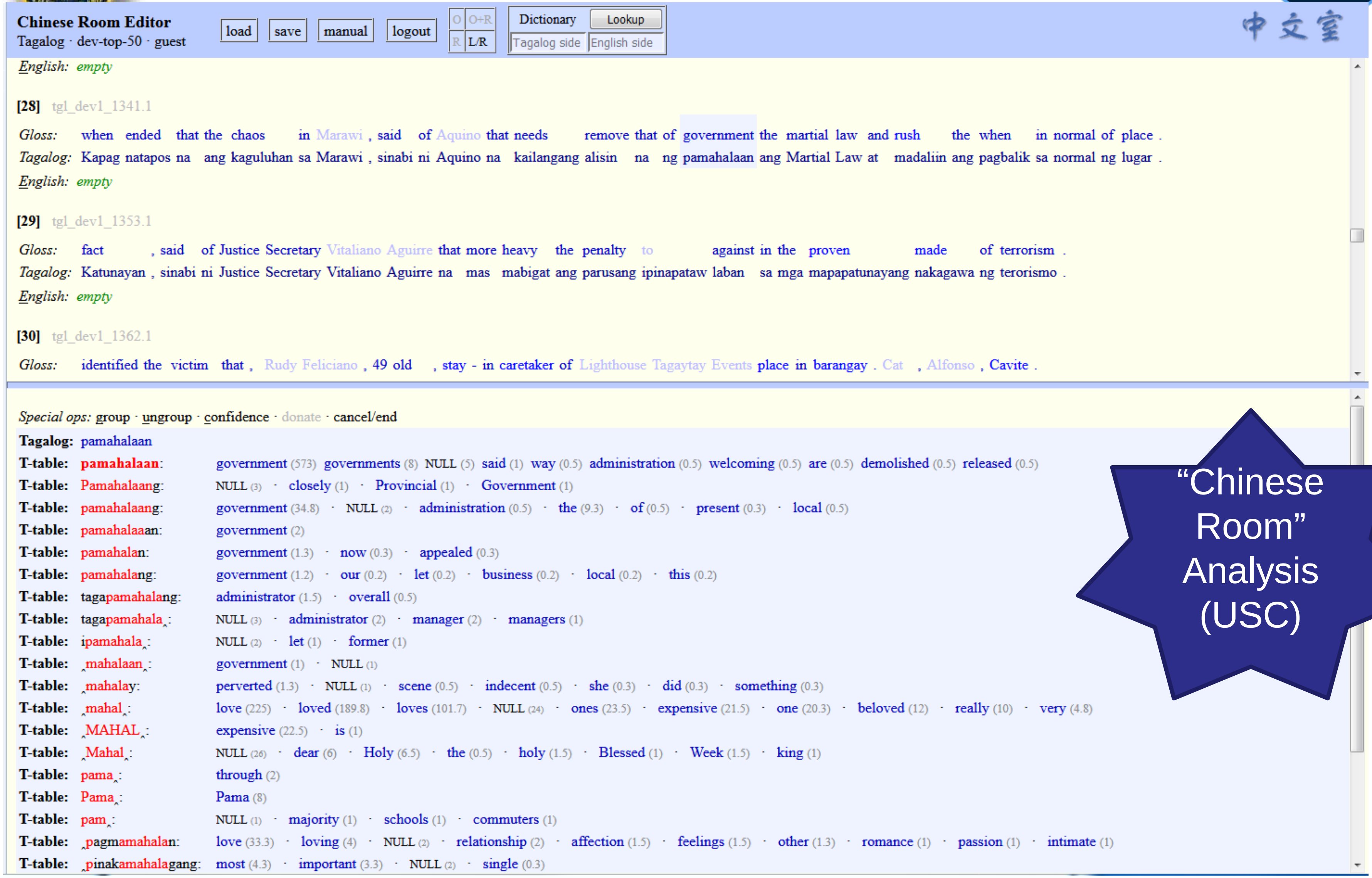
Adapted from:
Philipp Koehn
Johns Hopkins University

Substantial OOV Problem

- Languages with a large number of word variants will not be covered by the very limited amounts of training data.
- Doing well with limited training data requires accurate search for word forms not observed in training.
- 10% of queries OOV from English training; 15% not seen in Analysis Pack



Vocabulary Growth Rate – BABEL Languages



English: empty

[28] tgl_dev1_1341.1

Gloss: when ended that the chaos in Marawi , said of Aquino that needs remove that of government the martial law and rush the when in normal of place .
Tagalog: Kapag natapos na ang kaguluhan sa Marawi , sinabi ni Aquino na kailangang alisin na ng pamahalaan ang Martial Law at madaliin ang pagbalik sa normal ng lugar .
English: empty

[29] tgl_dev1_1353.1

Gloss: fact , said of Justice Secretary Vitaliano Aguirre that more heavy the penalty to against in the proven made of terrorism .
Tagalog: Katunayan , sinabi ni Justice Secretary Vitaliano Aguirre na mas mabigat ang parusang ipinapataw laban sa mga mapapatunayang nakagawa ng terorismo .
English: empty

[30] tgl_dev1_1362.1

Gloss: identified the victim that , Rudy Feliciano , 49 old , stay - in caretaker of Lighthouse Tagaytay Events place in barangay . Cat , Alfonso , Cavite .

Special ops: group · ungroup · confidence · donate · cancel/end

Tagalog: pamahalaan

T-table:	**pamahalaan**:	government (573) governments (8) NULL (5) said (1) way (0.5) administration (0.5) welcoming (0.5) are (0.5) demolished (0.5) released (0.5)
T-table:	Pamahalaang:	NULL (3) · closely (1) · Provincial (1) · Government (1)
T-table:	pamahalaang:	government (34.8) · NULL (2) · administration (0.5) · the (9.3) · of (0.5) · present (0.3) · local (0.5)
T-table:	pamahalaaan:	government (2)
T-table:	pamahalan:	government (1.3) · now (0.3) · appealed (0.3)
T-table:	pamahalang:	government (1.2) · our (0.2) · let (0.2) · business (0.2) · local (0.2) · this (0.2)
T-table:	tagapamahalang:	administrator (1.5) · overall (0.5)
T-table:	tagapamahala‸:	NULL (3) · administrator (2) · manager (2) · managers (1)
T-table:	ipamahala‸:	NULL (2) · let (1) · former (1)
T-table:	‸mahalaan‸:	government (1) · NULL (1)
T-table:	‸mahalay:	perverted (1.3) · NULL (1) · scene (0.5) · indecent (0.5) · she (0.3) · did (0.3) · something (0.3)
T-table:	‸mahal‸:	love (225) · loved (189.8) · loves (101.7) · NULL (24) · ones (23.5) · expensive (21.5) · one (20.3) · beloved (12) · really (10) · very (4.8)
T-table:	‸MAHAL‸:	expensive (22.5) · is (1)
T-table:	‸Mahal‸:	NULL (26) · dear (6) · Holy (6.5) · the (0.5) · holy (1.5) · Blessed (1) · Week (1.5) · king (1)
T-table:	pama‸:	through (2)
T-table:	Pama‸:	Pama (8)
T-table:	pam‸:	NULL (1) · majority (1) · schools (1) · commuters (1)
T-table:	‸pagmamahalan:	love (33.3) · loving (4) · NULL (2) · relationship (2) · affection (1.5) · feelings (1.5) · other (1.3) · romance (1) · passion (1) · intimate (1)
T-table:	‸pinakamahalagang:	most (4.3) · important (3.3) · NULL (2) · single (0.3)

“Chinese Room” Analysis (USC)

Data Crawling

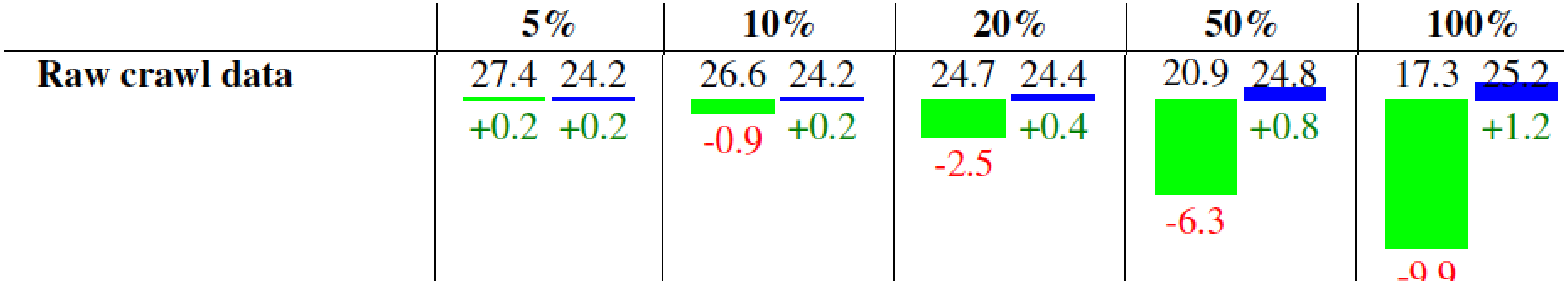
- Performers collecting heterogeneous monolingual and parallel corpora from large number of sources
  - **Text**: News/tabloids/blogs, Bible/Quran, bilingual dicts, Wikipedia
  - **Audio/Video**: News, YouTube, Bible

	Swahili	Tagalog
Text	• 3M documents, 130M sentences • 22K dictionary entries • 15K Bible verses	• 9M documents, 300M sentences • 36K dictionary entries • 63K Bible verses
Audio	• 93 hours of YouTube videos • 1600 hours of News videos • 99 hours of Audio Bible	• 5 hours of YouTube videos • 260 hours of News videos • 25 hours of Audio Bible

- Provided boost in ASR and MT domain/genre adaptations; eventually will be used in *all* MATERIAL tasks

Using Raw Crawled Data

- Unfiltered harvest, high-recall crawling output
- Data from Johns Hopkins University



	5%		10%		20%		50%		100%	
Raw crawl data	27.4	24.2	26.6	24.2	24.7	24.4	20.9	24.8	17.3	25.2
	+0.2	+0.2	-0.9	+0.2	-2.5	+0.4	-6.3	+0.8	-9.9	+1.2

Benefit to SMT 🟦, severe harm to NMT 🟩

Comparing MT Systems with BLEU

(Columbia's Baseline MT Results on Eval using Bitexts only)

MT Engine	Config	SW-EN	TL-EN	EN-SW	EN-TL
Neural (Marian)	Baseline	21.82	25.73	25.98	21.21
	Best (deep + ensemble)	31.33	**32.54**	36.00	30.51
Neural (Sockeye)	Baseline	30.16	28.83	36.01	27.49
	Best (tied embeddings+ ensemble)	**32.58**	32.10	**39.75**	**31.38**
Phrase Based (Moses)	Baseline	31.02	29.81	36.92	28.33

Comparing MT Systems with BLEU

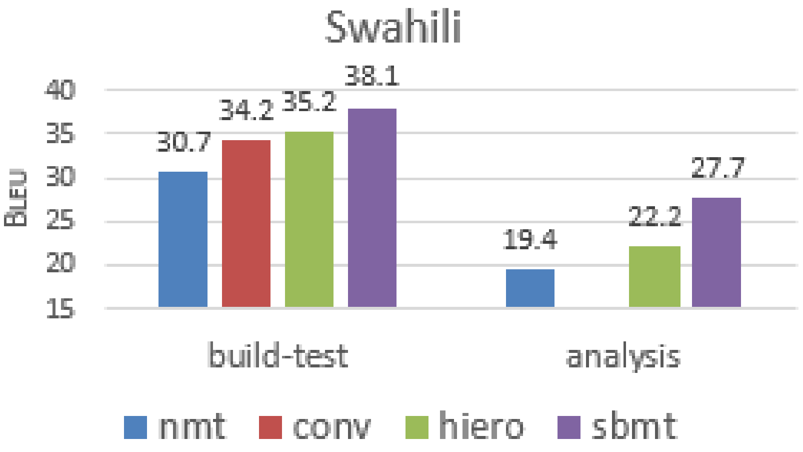
(Columbia's Improved MT Results using parallel + monolingual texts)

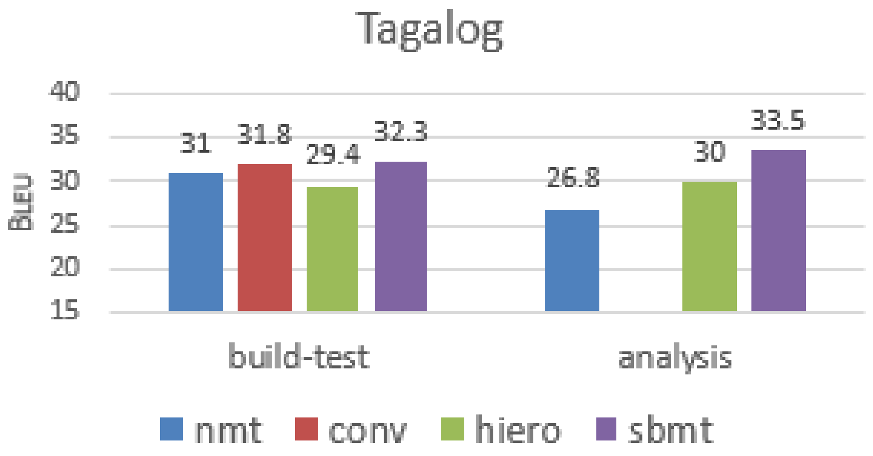
MT Engine	Config	SW-EN	TL-EN	EN-SW	EN-TL
Neural (Sockeye)	Parallel only	32.58	32.10	39.75	31.38
	+monolingual (best config)	**35.36**	**34.16**	**42.42**	**33.32**
Phrase-based (Moses)	Parallel only	31.02	29.81	36.92	28.33
	+monolingual (target LM)	31.94	30.78	41.57	30.24

Comparing MT Systems with BLEU

(ISI’s 90%/6.5%/3.5% train/dev/test split of the build set)

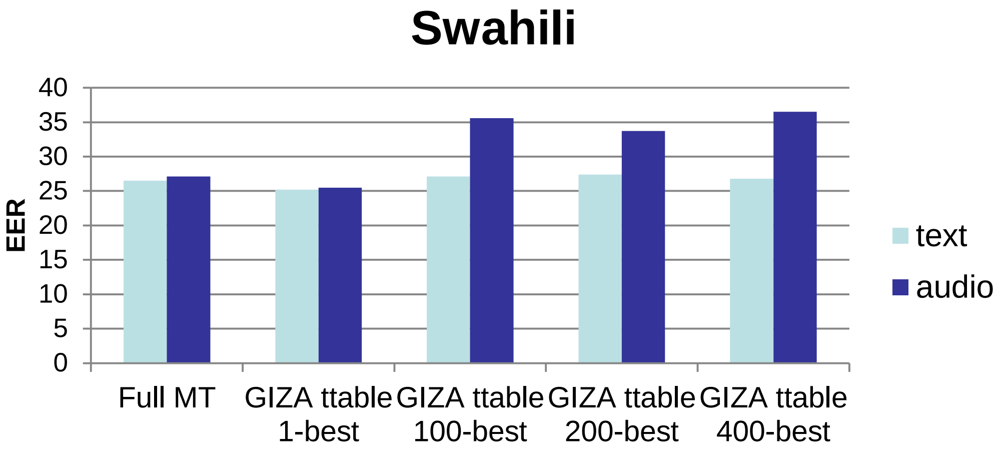
- Baseline systems
  - sbmt: Syntactic string-to-tree (from Galley et al. 2004)
  - hiero: Hierarchical Phrase-Based (from Chiang, 2005)
  - nmt: Neural seq2seq (Luo and Barzilay 2018, based on Bahdanau et al. 2014)
  - conv: Pure convolutional neural (adapted from Gehring et al., 2017)

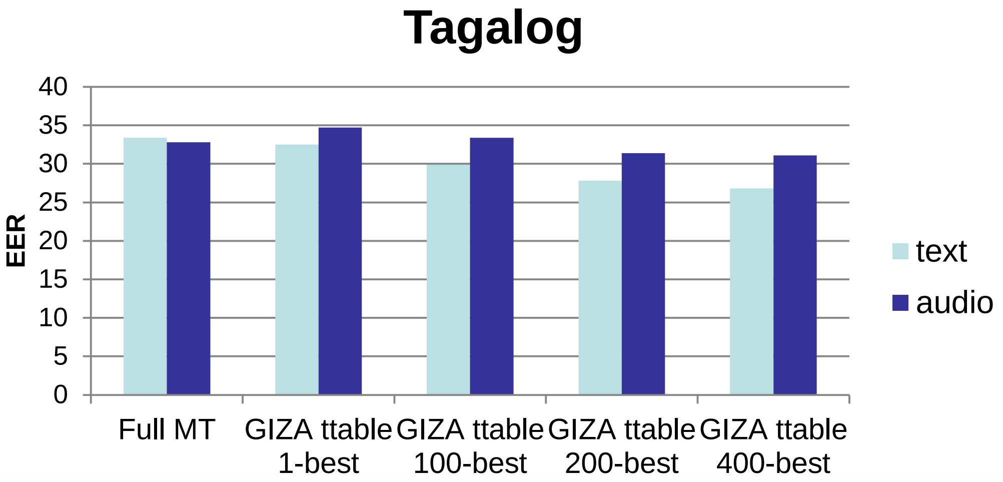




DOMAIN DETECTION

Equal Error Rate results with seeded Analysis 1 docs





Reminder: “GIZA ttable” is a translation table that contains word-to-word probabilities computed from the GIZA alignments of the parallel data.

The two practice languages behave quite differently!

Source: BBN

What’s Next?

MATERIAL BASE PERIOD SCHEDULE	
April 6, 2018	Development Cycle for Base Period Ends
May 14, 2018	CLIR Eval; Practice Summary Submissions
May 18, 2018	Domain 5 Released for both languages
May 25, 2018	CLIR/Domain ID Results Disseminated
July 5, 2018	Third query set release; Eval 3 Data released
July 24, 2018	CLIR+Summary Dry Run
Aug 6, 2018	CLIR+Summary Eval
Aug 14, 2018	Analysis Set 3 Release
Sept 5, 2018	Surprise Language Kickoff; DC Area Workshop
January 2019	CLIR+S Eval for Surprise Language
March 2019	CLIR+S Results Disseminated; Performer downselect

How to Engage with IARPA

Getting Started with IARPA

At IARPA, we take real risks, solve hard problems, and invest in high-risk/high-payoff research that has the potential to provide our nation with an overwhelming intelligence advantage.

Are you interested in partnering with us to advance the state-of-the-art in research and development?

iarpa.gov | 301-851-7500

info@iarpa.gov

Reach out to our Program Managers.

Schedule a visit if you are in the DC area
or invite us to visit you

Opportunities to Engage:

RFIS AND WORKSHOPS

Opportunities to learn what is coming, and to influence programs.

"SEEDLINGS"

Typically a 9-12 month study; you can submit your research proposal at any time. We strongly encourage informal discussion with a PM before proposal submission.

PRIZE CHALLENGES

No proposals required. Submit solutions to our problems – if your solutions are the best, you receive a cash prize and bragging rights.

RESEARCH PROGRAMS

Multi-year research funding opportunities on specific topics.

Questions?
May tanong ba?
Maswali?