Handling Entities in Machine Translation, Computer Assisted Translation, and Human Language Technology

Keith Miller, PhD

Linda Moreau, PhD

Sherri Condon, PhD

AMTA 2014

MITRE

Outline

- **Part 1: Name representation across languages, scripts, and cultures**
  - Why is entity translation important?
  - Survey of problems for entity translation
  - Transliteration
    - Transliteration standards
    - Automated transliteration
- **Part 2: State of the art and future directions for entity handling in MT/CAT**
  - Entities in isolation
    - Structured data
    - Unstructured data (search queries, extracted names)
  - Entities in context: MT/CAT
  - Evaluation approaches
  - Evaluation exercise

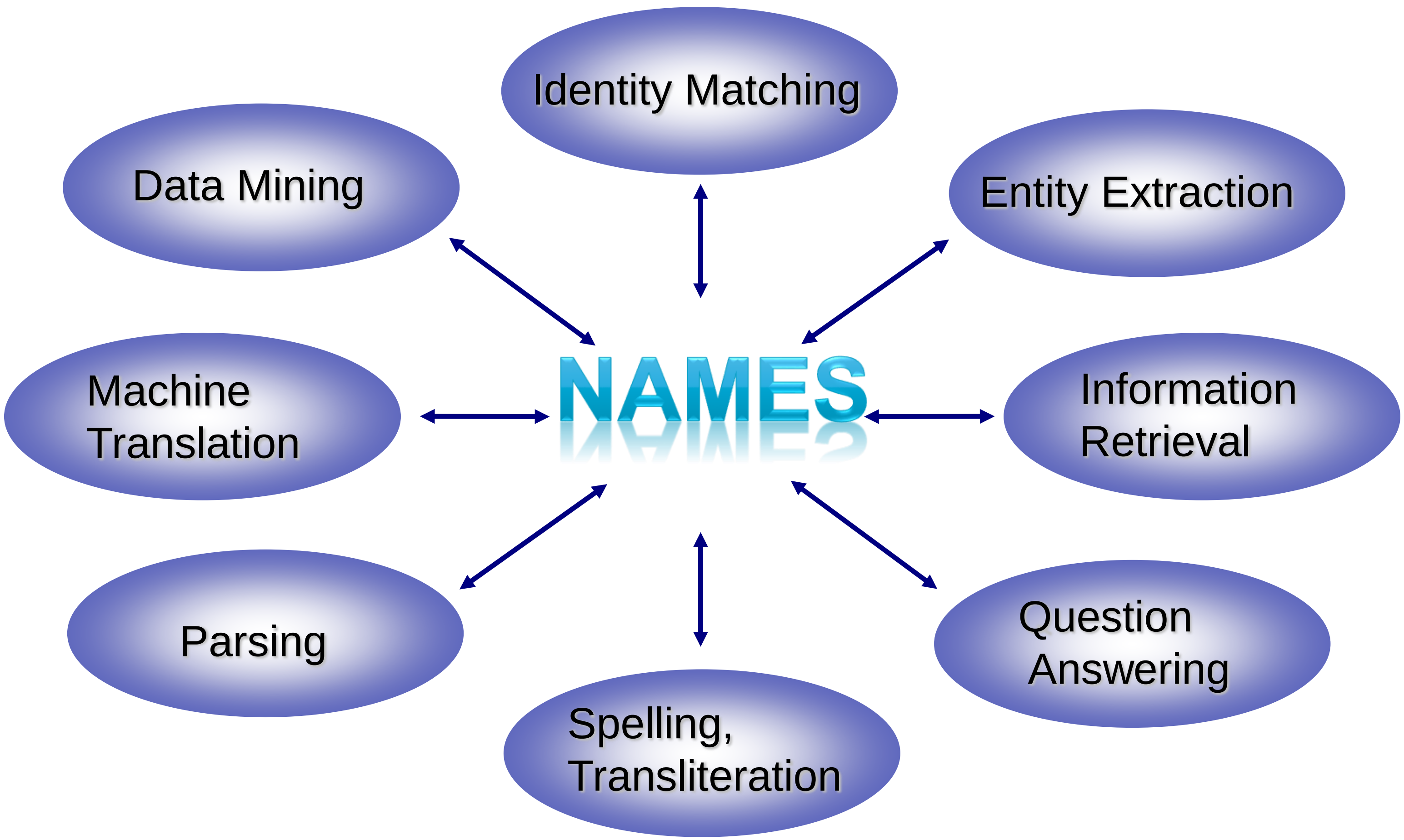
MITRE

Why is entity translation important?

Why is Entity Translation Important?

- **Information retrieval**
  - Entity names are typically key terms for embedded uses like Cross Language Information Retrieval (CLIR)
- **Structured data translation**
  - Data tables are typically focused on entity names and related data
- **Gisting and summarization**
  - Entities often represent the most significant information that is needed from a translated text: who, what, when, where...?
- **Automatic Translation**
  - Poor translation of entity names can cause poor translation of surrounding text

Impact: Embedded Uses of Entity Translation



Sources of Names in Computation

- **Written**
  - Hand print or script
  - Document images
  - Digital text
    - Prose / narrative
    - SMS
    - Email
    - Blogs
    - Structured data tables
- **Oral or oral-like sources**
  - Audio/video
  - Telephone
  - In person (mouth-to-ear)
  - Mental pronunciation / memory

MITRE

Survey of Problems for Entity Translation

First Activity

Morning Calisthenics!

NOTE: this exercise consists of transcribing 3 spoken names. The slide that discusses this exercise has been deleted from this version of the presentation. Possible answers to other exercises are also not included in this version.

Why is Entity Translation Hard? (#1)

- **Out-Of-Vocabulary (OOV) Problem**
  – Names are a rapidly expanding open class: they cannot be enumerated.
- **Data acquisition**
  – Noisy channels in written and oral transmissions of names add to the translation challenge.
- **Name detection**
  – Names are often homonyms or homographs of common nouns or adjectives. Poor translation of entity names can cause poor translation of surrounding text
- **Name-internal grammar**
  – Names are multi-word expressions that must be translated as a unit.

Why is Entity Translation Hard? (#2)

- **Differing cultural and linguistic conventions regarding names**
  - Each combination of language and entity type has unique features on most linguistic planes: phonological, orthographic, morphological and syntactic.
- **Transliteration challenges**
  - Transliteration is an inexact science due to imperfect alignments of phoneme and grapheme inventories.
- **Data exchange / data quality**
  - Data acquisition systems offer different data models between systems, and such models tend to reflect the naming conventions local to where the system is developed.
  - Standards for the exchange of name data are ill-defined or non-existent.
- **Idiosyncrasy**
  - In many languages, names have atypical phonological properties
  - They may preserve patterns not used in modern varieties
  - They are influenced by other languages and cultures

Second Activity: Segmentation

- **Which name segment is the family name?**
  - Anglo: Marianne Smith Miller
  - Hispanic: Maria Jose Gonzalez Hernandez
  - Arabic: Jaffar Abu Qasim Abd al Rahman

MITRE

Personal Name Challenges

- **Element variation**
  - Data errors
    - OCR
    - Typos
    - Truncations
  - Short forms
    - Abbreviations ***(Mhmd)***
    - Initials
  - Spelling variations
    - Alternate spellings ***(Karen, Karyn)***
    - Transliterations ***(Muhammad, Mohamed)***
  - Particles ***(von, de, bin, abu)***
    - Particle segmentation
    - Particle omission
  - Nicknames/diminutives ***(Bob,Joey)***
  - Translation variants
  - Non-word characters
  - Presence/absence of
    - Titles ***(COL, Dr., Ph.D.)***
    - Affixes***(-vich, -ovic, -ov)***
    - Qualifier ***(Jr., II)***
  - Case variation
- **Structural variation**
  - Additions/deletions
  - Fielding variations
  - Permutations
  - Placeholders
  - Element segmentation

MITRE

Other Cultures, Other Conventions

- **Different name segments carry different information value**
  - Most important segment of surname can vary according to cultural conventions
- **“Phases of life” can influence name used**
  - Haj/Haji, Vda/V de, married name, confirmation name, Dr.
- **Importance placed on given name varies**
  - Common practice of using familiar name / nickname
- **Frequency of surnames / given names varies**
  - e.g. Smith; Korean family names; Muhammed
- **Romanization from different scripts introduces other challenges**
- **May have completely different naming model**
- **Complication for ID matching in general:**
  - Lack of emphasis on record keeping: e.g. inexact or unavailable birth dates

MITRE

Arabic Example: Name Variants

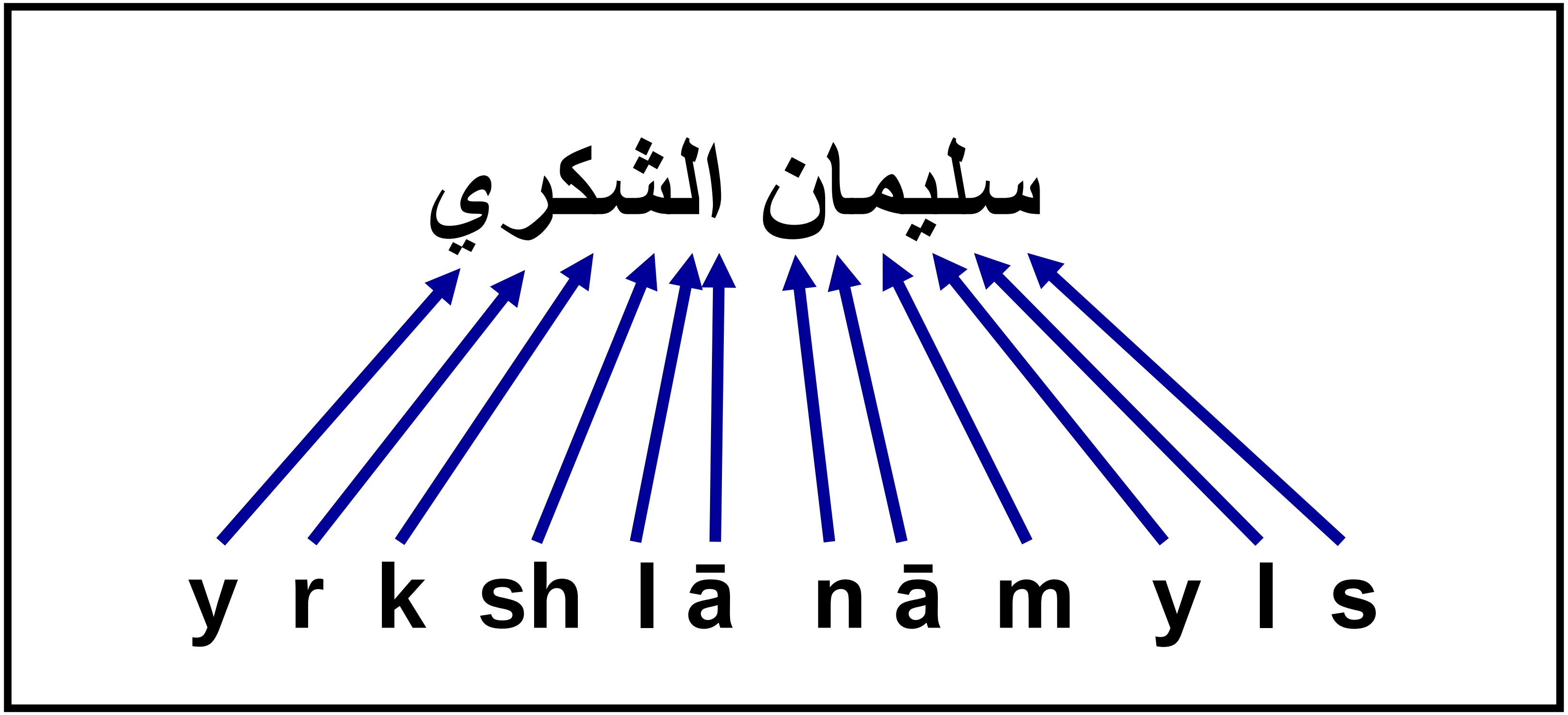
سليمان الشكري

Sulayman al-Shukri	Soleiman Shukri	Suleyman Shukri
Solomon Ash-shukri	Sulejman Ashukri	Suleman Schoukri
Suleiman Alshokri	Suleman Al-Shukri	Soleiman Choukri
Süleyman Alshukri	Soulaiman Choukri	Soulaiman Achoukri
Sulejman Shukri	Suleman Shukri	Süleyman Shukri
Suleman al-Schoukri	Soloman Ash-shukri	Suliman Al Shukri
Soleiman Ashukri	Solomon Shukri	Soulaiman Al Choukri
Soulaiman al-Choukri	Suleyman Alshukri	Sulejman Ashukri

سليمان محمد حسين الشكري

Sulayman Muhammad Husayn al-Shukri

Arabic Example : Why all that variation?



ع عع → ‘ , a, other vowel, or deleted

One-to-many and many-to-one mappings

Arabic Example: Phoneme Inventories

	bilabial	labio-dental	inter-dental	dental	alveolar	postalveolar	retroflex	palatal	velar	uvular	pharyngeal	glottal
nasal	m				n				ŋ			
plosive +voice	b			dˤ d̪	d				g			
plosive -voice	p p^h			tˤ t̪	t^h t				k k^h	q		ʔ
fricative -voice		f	θ		sˤ s	ʃ			x		ħ	h
fricative +voice		v	ð ðˤ		z	ʒ			ɣ		ʕ	
approximant					ɹ			j	w			
trill/tap/flap.					r ɾ							
lateral approximant					l							

▪ English ◈ Arabic

MITRE

Arabic Example : Personal Name Structure

- **Given name**
- **Father's given name**
- **Grandfather's given name**
- **Family name**
- **A geographic or tribal name, which is usually preceded by *al* "the" and followed by the suffix –*i*, e.g. *al Basri* "from Basr."**

❖ **Note:**

The patronymic (fathers') names may or may not be preceded by *bin* "son of"

The given name may also include a descriptive name, usually religious, such as *'Abd Allah* "Servant of God" (often written *Abdullah)* or with abu "father of"

MITRE

Arabic Example: Data Capture

FullNameString: Maria Hernandez de Rodriguez

NameFormat: DerivedNameFormat

NameCategory: ProvidedName

DerivedNameInfo

- DerivedFromField: FullNameString
- DerivedFromName: ماريا ايرناندز دي رودريغز
- DerivedType: ICArabicTransliteration

ParsedName

ParsedName

- SurnameString: Hernandez de Rodriguez
  maxlength: 50
- GivenNameString: Maria
  maxlength: 50

NameParts

Name Part
Maria

Name Part
Hernandez

Name Part
De

Name Part
Rodriguez

NameScript: RomanScript

DerivedNameInfo

Data Exchange Formats: Name Object

Data capture and sharing can be challenging when name models used in capture systems differ from the conventions of other cultures

MITRE

Arabic Example: Transliteration

Transliteration introduces more dimensions of variation

Issue	Example
Multiple standards	BGN, LOC, IC, Buckwalter, SATTS, …
Multiple traditions	Francophone tradition (Wasim = Ouassime)
Acoustic errors	Ali = 'Ali
Dialectical variants	Bourguiba = Abu Ruqayba
Non-native names / N-way transliteration	Pavel = Bafil
Segmentation	Abd Al Rahman = 'Abdurrahman
N-to-n mappings	Walid = وليد and واله
Missing information	محمد = mhmd

MITRE

Location Name Challenges (#1)

- **Mix of translation and transliteration**
  - гора Кошка⇔Mount Koshka *not* Mount Cat
- **Morphology**
  - О́мская о́бласть ⇔ Omsk Oblast
- **Reverse transliteration**
  - ボストン /bosuton/⇔ Boston
- **Absent name parts**
  - the Mississippi *vs.* the Mississippi River
- **variants**
  - The United States of America, the USA, U.S., E.E.U.U.
- **nicknames**
  - The Windy City, The Big Apple

Location Name Challenges (#2)

- **Domain and category dependent word sense disambiguation**
  – Mesa Central
- **Abbreviations**
  – Mt., Rte. , ул., г., Str., St. (Saint or Street?)
- **Country-specific administrative divisions**
  – Oblast, Prefectura, Länder
- **Geographic feature ontology differences**
  – river ⬄ fleuve/rivière
- **Idiosyncratic translations**
  – Bahía de Fundy⬄ Bay of Fundy *vs.* Bahía de Hudson⬄ Hudson Bay
- **Multi-token morphology/syntax**
  – Little Harbor on the Hillsboro, FL

Organization Name Challenges (#1)

- **Mix of translation and transliteration**
  - 삼육대학교 ⇔ Sahmyook University
- **Morphology**
  - Ивáновский госудáрственный университéт ⇔ Ivanovo State University
- **Reverse transliteration**
  - دانشکده پزشکی آلبرت اینشتین ⇔ Albert Einstein College of Medicine
- **Compounds and portmanteaus**
  - *Bricomarché, Artbambou, Brico-Depôt*
- **Absent name parts**
  - Carrefour, Groupe Carrefour, Carrefour, S.A.

Organization Name Challenges (#2)

- **Variants, *long/short/legal/informal forms***
  - NYS Dept. of Energy ⇔Energy Department of New York State
- **Variants, *nicknames***
  - Wally World, The Evil Empire
- **Complex syntax and embedded entities**
  - Musée d’art et d’archéologie de l’Université d’Antananarivo à Tananarive
- **Domain and category dependent word sense disambiguation**
  - la **Mesa** del CIG⇔IGC **Bureau** (ORG) *vs.*
  - Tienda de **Mesas** de Billar⇔Pool **Table** Shop (ORG) *vs.*
  - **Mesa** de Wingate ⇔ Wingate **Mesa** (LOC) *vs.*
  - Alfredo **Mesa** ⇔ Alfredo **Mesa** (PERS)

Organization Name Challenges (#3)

- **Abbreviations**
  - Dept., Grp. Cntr.
- **Organizational legal ontology differences**
  - SàRL, Inc., GmBH
- **Preferred syntax**
  - Auto-école Conduite Sans Frontières ⇔ Without Borders Driving School (*probably not* Driving School Driving Without Borders)

MITRE

Summary of Named Entity Challenges

	PERSON	LOCATION	ORGANIZATION
Abbreviations	X (Initials)	X(esp. of keywords)	X (esp. of keywords)
Short forms	X (nicknames, diminutives)	X(e.g. full legal, short common)	X (acronyms, no org designator)
Variants	X (esp. transliterated and nicknames)	X (e.g. local names)	X (nicknames, branch names)
Mixed translation/transliteration	X (titles, qualifiers)	X	X
Entity-specific morphology	X (e.g. qualifiers, patronymic suffixes, name particles)	X (location suffixes, prepositions)	X (novel compounds, portmanteaus)
Inflection of names in context	X	X	X
Absent name parts	x	x	x
Incorrect fielding	X	X	X
Reverse transliteration	X	X	X
Entity-specific syntax	X	X	X
Domain- and category-dependent senses	X	X	X
Cross-language ontology issues	X (titles, honorifics, degrees)	X (e.g. lagoon, pond, sea and admin levels)	X (e.g. untranslatable org designations)
Idiosyncratic word ordering		X (local/historical convention)	X

Transliteration

Used Here *Transliteration* is Not:

- **Transcription:**
  – Renders speech sounds into written characters
- **Character mapping:**
  – Associates each character in a set of characters with a character in another set of characters
    - Usually without regard to context or meaning
    - Possibly without regard to pronunciation
    - Emphasis on consistency
    - Usually reversible/lossless/one-to-one
    - Example: محمد = mHmd (vs. Muhammad)

Transliteration

- Renders written words from one language into the written forms of another language in a way that reflects the sounds and/or spellings of the original, rather than the meaning
- Usually names of people, places and organizations
- May incorporate special conventions for context or function
- Usually tries to reflect pronunciation
- Often sacrifices reversibility for readability

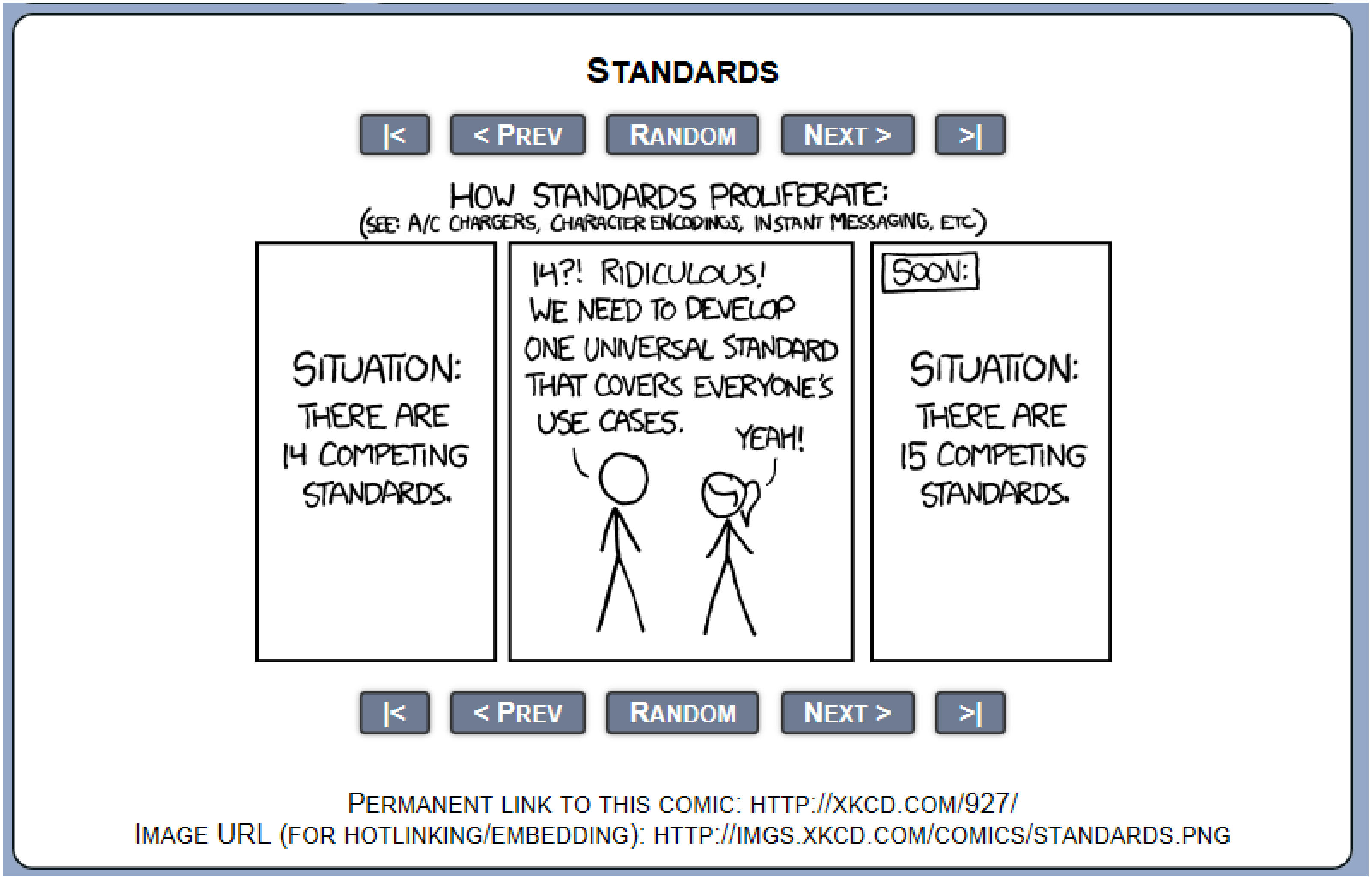
Transliteration Standards

- **Transliteration standards specify mappings for transliteration**
- **The goal is to eliminate transliteration variants by providing consistent mappings**
- **But this goal has not been achieved**
  - Failure to apply standards: people make up their own spellings
  - Errors in applying standards
  - Multiple standards

Arabic Standards	Chinese Standards
Board of Geographic Names (BGN)	BGN
Intelligence Community (IC) Standard for Person Names	IC Standard
Buckwalter	Hanyu Pinyin
SATTS	Wade-Giles

MITRE

By Whose Standard?



Why Multiple Transliteration Standards?

- **Different transliteration systems satisfy different constraints and goals**
  - One-to-one mapping, which makes the transliteration reversible and lossless
  - Readability
  - “Type-ability”
  - No distinctions between upper and lower case letters (for State Department cables, which are all caps)
  - No digraphs (though English already has *th, sh, ch*)
- **Some constraints and goals are mutually exclusive, e.g., one-to-one mapping and readability in Arabic *(mhmd* vs. *Muhammad)***
- **Governments may impose standards (Pinyin, BGN, IC Standard)**

MITRE

Transliteration Types

- **Forward transliteration**
  - Conversion from the native form of a word in the original language to the transliterated form in another language.
- **Backward transliteration**
  - Conversion from the transliterated form of a word in one language to its native form in the original language.
- **N-Way transliteration**
  - In many contexts these two types are incomplete because additional languages are involved, e.g. transliterating a Chinese name from Arabic into English.

Transliteration Challenges

- **Preprocessing sometimes necessary**
  - Orthographic reasons
    - Semitic languages & vocalization
      - Rule based, statistical, dictionary based
  - Phonotactic
    - Japanese, Chinese syllable structure

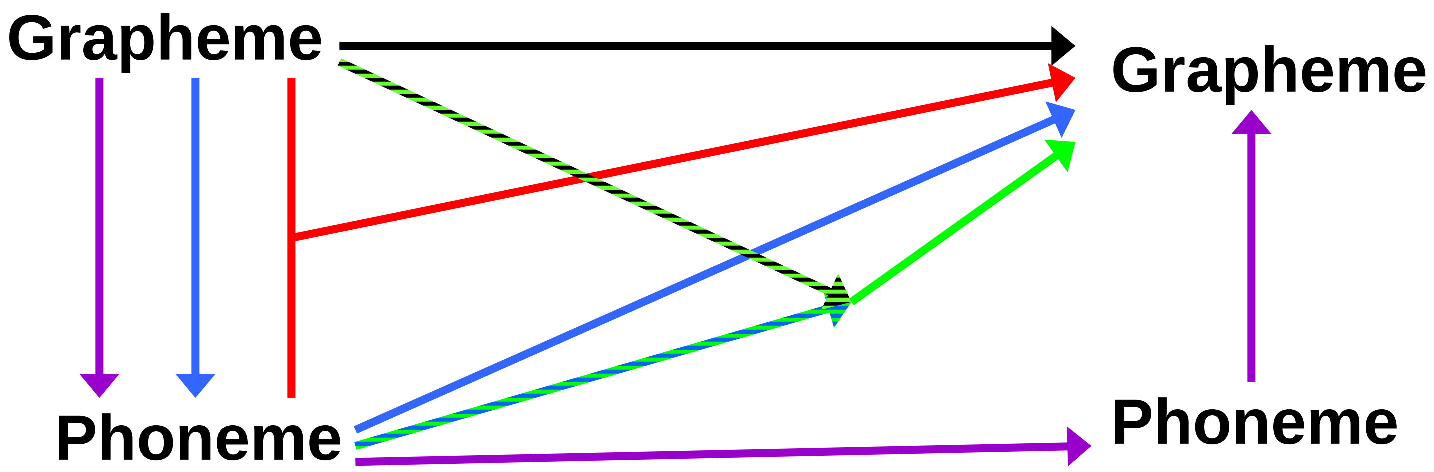
Rank (2007)	Trad.	Simp.	Pinyin	Wade-Giles
52	盧	卢	Lú	Lu^2
47	呂	吕	Lǚ	$Lü^3$
57	陸	陆	Lù	Lu^4

Derived from 08/31/2014 version of http://en.wikipedia.org/wiki/List_of_common_Chinese_surnames,

- **Conversion can be lossy / destructive**
  - Many-to-one conversions
    - 'r' and 'l' → ラ (Katakana 'ra')
  - One-to-many conversions
    - 's' → 'س' or 'ص'
  - Phonetically required insertions alter syllable structure
    - オペレイティングシステム ：(Opereitingu shisutemu)
    - コンピューータープログラマー：（Konpyuutaa Puroguramaa)
    - イングランド ：（Ingurando)
    - シンドローム：（Shindoroomu ）
  - Tone often ignored
    - Chinese/Thai -> English

MITRE

Automatic Transliteration Choices



1. Grapheme to grapheme
2. Grapheme to phoneme to grapheme
3. Grapheme+phoneme correspondence to grapheme
4. Grapheme to grapheme and phoneme to grapheme hybrid
5. Grapheme to phoneme to phoneme to grapheme

Grapheme to Grapheme
Example: Al-Onaizan & Knight (2002a)

- **For word sequence *w, P(w)* is a unigram model that generates English word sequences according to their unigram probability**
  - Estimated from word lists (*Wall Street Journal*, names)
- **Transliteration maximizes $P_s(w|a) \simeq P(w)\,P(a|w)$, *a is an Arabic sequence***
- ***P(a|w)* is estimated from English - Arabic pairs**
  - Estimate symbol mapping probabilities using Estimation Maximization for values in a WFST
  - 1 – 3 English letters are mapped to 0-2 Arabic graphemes
  - Incorporates position: initial, medial, final

mahar g (Graham)

غراهام (= /ġrāhām/)

Note: the formulas above are for Arabic to English transliteration, but the example is English to Arabic in order to illustrate the consequences of the unigram model

MITRE

Grapheme → Phoneme → Grapheme Example: Al-Onaizan & Knight (2002a)

- **For English word sequence *w* and English phoneme sequence *e***

$$P_p\,(w|a) \simeq \sum_{\forall e} P(w)\,P(e|w)\,P(a|e)$$

- ***P(e|w)* is estimated from CMU pronouncing dictionary**
- ***P(a|e)* is estimated from 1426 English – Arabic name pairs**
  - Positions are handled using 3 states for initial, medial, and final
  - Each English phoneme maps to 0 or more Arabic graphemes
  - Transliteration is a graph search to maximize *P(w|a)*

Graham →/gram/ ⇨ /ma rg/

↓↓↓↓

غرام

Note: the formulas above are for Arabic to English transliteration, but the example is English to Arabic in order to contrast with the example on the previous slide

MITRE

Grapheme+Phoneme to Grapheme Correspondence

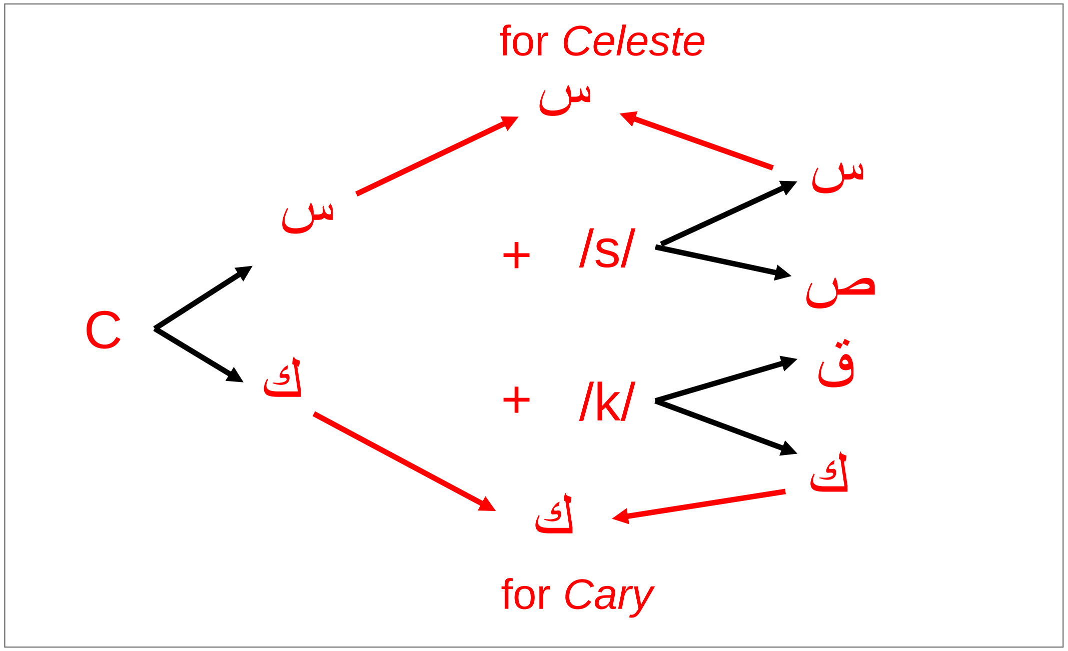
- **Example: Oh & Choi (2002, 2006)**
- **Grapheme – phoneme correspondence in L1 maps to grapheme in L2**
  - Context sensitive rules for English to Korean transliteration
  - English grapheme *r* corresponding to English phoneme /r/ maps to null in Korean following vowels

Cary	Celeste
/ k a r i /	/ s ə l ɛ s t /
ك ا ر ى	س ا ل ا س ت

MITRE

Grapheme+Phoneme to Grapheme Hybrid

- **Grapheme – grapheme and phoneme – grapheme probabilities are combined**
  - Example Onaizan & Knight (2002b)
  - $P(w|a) = \lambda P_s(w|a) + (1 - \lambda) P_p(w|a)$



MITRE

Grapheme → Phoneme → Phoneme → Grapheme
Example: Knight & Graehl (1997)

- ***P(w)*** **WFSA for English word sequences**
- ***P(e|w)*** **WFST maps to English phonemes**
- ***P(j|e)*** **WFST maps to Japanese phonemes**
  - Estimation maximization to learn alignment probabilities
- ***P(k|j)*** **WFST maps to katakana**
- **Maximizes the sum over all e, j, and k of**

$$P(w) \cdot P(e|w) \cdot P(j|e) \cdot P(k|j)$$

علي → /ʕali/ → /ali/ → Aly? Ally? Allie?

MITRE

Transliteration Choice Comparison

L	Advantage	G⇨G	G→P⇨G	G+P⇨G	G→P⇨P→G
B	Directly models grapheme correspondences	✓	✗	✓	✗
B	Directly models phoneme correspondences	✗	✗	✗	✓
S	Addresses effect of irregular spelling	✗	✓	✓	✓
T	Addresses effect of irregular spelling	✗	✗	✗	✓
S	Addresses effect of pronunciation variation	✓	✗	✓	✗
T	Addresses effect of pronunciation variation	✓	✗	✗	✗
S	Avoids mapping of graphemes to phonemes	✓	✗	✗	✗
T	Avoids mapping of phonemes to graphemes	✓	✓	✓	✗

L = language B=both S=source T=target G=grapheme P=phoneme

Variations

- **Handcrafted mappings**
  - Oh and Choi (2002) context sensitive rules were handcrafted
  - Wan & Verspoor (1998) fully handcrafted and rule-based mappings for English to Chinese Pinyin
  - Meng et al. (2001) handcrafted phonological normalization of English for transformation error-based learning of mapping into Chinese Pinyin
  - Jung, Hong & Paek (2000) handcrafted mapping between English and Korean phoneme pairs
- **Context**
  - Oh & Choi (2006) tested window size of 1 - 5
  - Jung, Hong & Paek (2000) used ±1 English phonemes and -1 Korean grapheme

Problems

- **Alignment**
- **Allowing segments to map to zero segments**
  - Expensive to compute
  - Huge numbers of hypotheses in WFST composition
  - Knight & Graehl (1997) prohibit this and removed hundreds of “harmful” pairs from the English-Japanese training set, which then require dictionary look-up
- **Errors can cascade**
- **Chinese many to many mappings**
  - Li, Zhang & Su (2004) joint source channel model

Chinese Pinyin Mappings

Number of distinct mappings	Chinese characters mapped to Pinyin forms	Pinyin forms mapped to Chinese characters
1	5708	260
2	753	168
3	111	151
4	17	114
5	5	104
6	1	76
7	1	64
>7	0	365

Based on calculations from LDC Corpus # LDC2003E07

MITRE

Web-Frequencies to Rank Candidates

- Oh & Choi (2005) and Al-Onaizan & Knight (2002b, 2009) use normalized Web counts to rescore transliteration candidates
- Onaizan & Knight (2002b) also use contextual web counts: name plus title or key words or local terms
- Huang (2004) uses TF-IDF to find similar documents and compares candidate translations using a transliteration similarity measure and a vector of context features (words and parts of speech)
- Jiang et al. (2007) search web with source name to find target terms similar to candidates, then search again with source name and higher scoring candidates and use top 30 texts returned to rank candidates using maximum entropy with features based on the number of web pages containing the terms

Web-Based Transliteration

- **Sproat, Tao & Zhai (2006); Tao et al. (2006)**
  - Identify candidate transliterations using comparable corpora, e.g. news articles about the same event in two different languages
  - Score candidates based on phonetic similarity and a frequency profile
  - Combine similarity and frequency scores
- **Oh & Choi (2005) search for source/transliteration pairs as phrases or in the same document (for chemical names)**
- **You et al. (2012) use entity search engines in English and Chinese to identify entity names and their co-occurrences with other entity names in documents on the Web**
  - A graph structure represents relations among the names separately in each language based on co-occurrence frequency
  - A similarity measure associates English and Chinese Pinyin name pairs for an initial match across the two languages, which is then optimized to match the names in each language that have the most similar graph structures

Transliteration Evaluation Issues

- **What is the "correct" transliteration?**
  - Frequently more than one transliteration is acceptable
  - Match scores computed against training data with a single transliteration will underestimate accuracy
  - Including more than one correct transliteration complicates computation of evaluation scores
- **Scores will vary according to data type, e.g. personal names vs. chemicals**
- **Human transliteration is frequently inaccurate**
  - Names may not be recognizable
  - ال غور al qur al gur Al Gore

Evaluation Measures

- **Edit distance**
  - Divide edit distance by length of transliteration
  - Three English to Chinese methods achieved about .5
- **Accuracy: exact match to gold standard**
  - Knight & Graehl (1997) 64% vs. 27% for humans
  - Onaizan & Knight (2002b) 72.57% with web counts
- **Recall, precision, and F score**
- **Error rates**
  - Character: Li, Zhang & Su (2004) report 10.8% CER for top choice in English to Chinese, 19.6% for Chinese to English
  - Word: Li, Zhang & Su E to C WER is 29.9%; C to E WER is 62.1%
- **Compare to Google translate (You et al. 2012)**
  - F is 0.74 vs. Google 0.75 for high frequency names
  - F is 0.69 vs. Google 0.56 for low frequency names

MITRE

Presentation of Measures

- **Training vs. test sets**
  - Most use cross fold validation
  - Sizes vary enormously
- **In dictionary vs. not in dictionary (for grapheme to phoneme mappings)**
- **N-best results**
  - Jung, Hong & Paek (2000) .875 recall for 10 best
  - Li, Zhang & Su (2004) E to C WER decreases to 5.4% and C to E WER decreases to 24.6% for 10 best
  - Mean Reciprocal Rank (MRR) Kantor & Voorhies (2000)

Resolve Variation with Matching

- **Obtaining one of many existing variants may not be adequate for downstream search and retrieval applications**
- **Satisfactory results are achieved by "fuzzy" matching instead of exact matching**
- **Matching techniques can be customized for specific languages**
- **Similar approaches can be used for matching across languages and scripts**

translit	score	freq
Gadhafi	1.0	21,300
Gadhaffi	0.975	83
Gadafi	0.966	2,330
Ghadafi	0.957	1,020
Gaddafi	0.933	17,000
Gadaffi	0.933	2,270
Ghadaffi	0.919	435
Ghadhafi	0.919	94
Khadafy	0.742	1,700
Kadaffy	0.714	52
Quadafy	0.714	43
Qaddafy	0.714	40
Khadaffy	0.713	797
Khaddafy	0.713	329
Khaddafy	0.713	285

Jaro-Winkler similarity scores for 'Gadhafi'

MITRE

Entities in Isolation

Entities in Isolation: Structured Data

- **Spreadsheets, CSV files, Database tables**
  - Entity data and supporting attributes
- **Issues**
  - CONTEXT: Sentence- or phrase-level context absent (some types of word-sense disambiguation more difficult or impossible). Categorization by column or entity type can help.
  - COMPLEXITY: Location and organization names are especially complex, and often have other embedded entity types in them (Person, Location, Organization Names)
  - VARIABILITY: Even in spreadsheets, values are not always constrained or predictable (e.g. Address could be just street level information or could be entire contact card including name; extraneous information can be included)

Structured Data: Sample Column Headers

PERSON	LOCATION	ORGANIZATION	RELATED CATEGORIES
Name	Address	Name	Gender
First Name	Street	Industry	Marital Status
Last Name	City	Company	Age
Complete Name	Region	Organization	Education
Maiden Name	Country	Enterprise	Industry
Alias	Nationality	Business	Occupation
Recipient	County	Partner	Religion
Addressee	Birthplace	Manufacturer	Ethnicity
Beneficiary	Origin	Employer	Relationship
Manager	Location	Institution	…
Contact	Headquarters	Recipient	
…	…	…	

MITRE

Structured Data Example

Company	ООО Компания Эриксон	ООО Алтайрегион Торговый Дом	ЗАО АПОСТРОФ ПРИНТ
Address	620016, Россия, г.Екатеринбург, ул.Амундсена 133, 2-ой этаж	656023,Россия,Барнаул,А/Я 4512.	117105, Россия, Москва, Варшавское шоссе, д. 37а
City	Екатеринбург	Барнаул	Г. Москва
Country	РОССИЯ	РО	РОССИЯ
Phone	+ 7 (343) 267-83-91	+ 7 (3852) 34-56-31, 33-02-37	
URL	*www.**erickson.ru***		http://www.apostrof-print.ru/
Contact	alex@erickson.ru	Исаева Татьяна Николаевна	+ 7 495 781-38-38
Position	Заместитель руководителя	Топ-менеджер по региональным продажам	

Entities in Isolation: Extracted Entities

- **Issues for entity data extracted from unstructured text**
  - EXTENT: Match could contain extra or missed spans of text
  - TYPE: Extracted entity type could be wrong
  - NONENTITIES: Extracted entities could be false positives
  - CONTENT: Inclusion of certain information, e.g. titles, dependent upon extraction algorithm
  - MORPHOLOGY: Inflectional morphology likely to be an issue (for inflected languages)

MITRE

Third Activity: Entity Categorization

- **Indicate whether each name is Person, Location , Organization or Other:**

Easy Street	
Benjamin Moore	
Clarion Alley	
T.S. Cooper	
T.S. Elliot	
Lively	
Christian Dior	
Honda	
Geneva Parks	
United Way	
Summer Lane	
Dom DeLuise	
Dom Perignon	
Miss Georgia	
Mayor Street	

Structured MT: Keyword Categorization

- **Knowing entity types may help produce better translations**
- **Categorization can be challenging based on presence of keywords alone, instead, a language's noun and/or adjective phrase headedness may be required to disambiguate**

Market Street Grille
United Way Foundation
Lee Jackson Memorial Highway
University Boulevard
Ronald Reagan Washington National Airport
Business Center Drive
Site Drive Inc.
Windshield Dr., Inc.
Duke Ellington School of the Arts
Mayor John F. Street
King Abedulla II Industrial City

MITRE

Structured MT: Abbreviations

- Expansion and or translation can be dependent upon:
  - Category
    - **St.** ⇔ Street *vs.* **St.** ⇔ Saint
    - **Dr.** ⇔ Drive *vs.* **Dr.** ⇔ Doctor
    - **г.** ⇔ город *vs.* **г.** ⇔ господин
  - Syntactic position
    - 265 **St.** Vincent **St.** Church
    - **м.** Братисловская, **ул.** Братиславска **д.** 10
    - **г.** Ижевск, **ул С.** Ковалевской, **д.** 12, **к.** 21
    - **U St**
    - **U St** Paul
  - Domain within category:

	Str.
International » German	Straße
Medical » Physiology	Straight
Governmental » Military	Strength
Medical » Physiology	Strength
Medical » Physiology	Strain

http://www.abbreviations.com/STR 09/03/2014

MITRE

Fourth Activity: Acronyms & Initialisms

- List possible expansions of the following acronyms in an ORG name:

EMT	AMS

MITRE

Structured MT: Output Normalization

- **Normalized or standardized forms for translated entities allow**
  - Support for database indexing
  - Increased retrieval for IR or CLIR applications
  - Support for entity clustering and co-reference applications

- **Example**
  - US, USA, United States, the United States, the United States of America, EEUU, can all be mapped to a single form

    - E.g. Virtus MT engine for structured data allows users to specify whether to output a standardized form for entities listed in the user terminology list and to update user terminologies to specify custom standard forms

Structured MT: Transliteration Standard Support in Mixed Names

- **Consistent output**
  - Transliterated portions of names in structured data should be transliterated according to a consistent scheme.
  - Entities retrieved from terminologies should be subject to the same scheme as algorithmically translated entities

Хакасский государственный университет имени **Никола**я **Федорович**а **Катанов**а

Khakas State University “named-after” **Nikolai Fedorovich Katanov**

عمان - مرج الحمام - مجمع النابلسي التجاري

Amman - **Marj Al Hamam** - **Al Nabilsi** Commercial Complex

MITRE

MT of Extracted Entities: Inflection

- **Inflected forms of entities need to be detected and translated**
- **Output required depends on language pair and intended use**
- **E.g.:**

Russian Adj-noun phrases agree in Number, Case, and Gender. The adjective takes on the value of Number, Case,and Gender from the head noun.

Московская область (Moscow Oblast)

Nominative:	Московская область
	Moskovskaia (ADJ:Nom.Fem.Sg.) oblastj (NOUN: Nom.Fem.Sg.)
Genitive:	Московской области
	Moskovskoj (ADJ:Gen.Fem.Sg.) oblasti (NOUN: Gen.Fem.Sg.)
Accusative:	Московскую область
	Moskovskuiu (ADJ:Acc.Fem.Sg.) oblastj (NOUN: Acc.Fem.Sg.)

Московский комбинат (Moscow factory)

Nominative:	Московский комбинат
	Moskovskij (ADJ:Nom.Masc.Sg.) kombinat* (NOUN: Nom.Masc.Sg.)
Genitive:	Московского комбината
	Moskovskogo (ADJ:Gen.Masc.Sg.) kombinata (NOUN: Gen.Masc.Sg.)
Dative:	Московскому комбинату
	Moskovskomu (ADJ:Dat.Masc.Sg.) kombinatu (NOUN: Dat.Masc.Sg.)

MITRE

MT of Entities: Stopwords

- **When matching against translation memories or lexical resources, some entity types may require selective stopword lists**

SEARCH TERM: "Physicians **For** Euthanasia"

TM ENTRY:

EN: "Physicians **Against** Euthanasia"

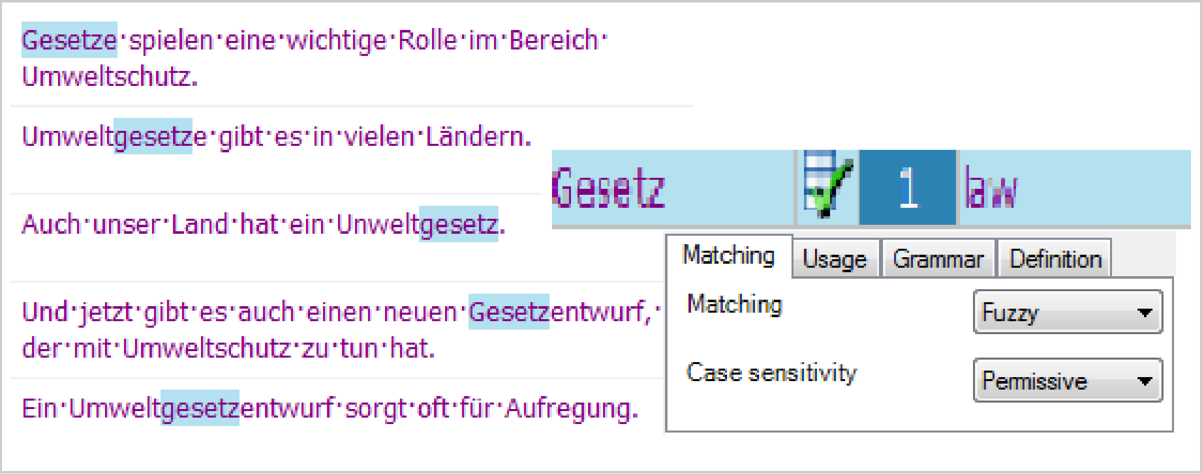
SP: "Médicos **contra** la eutanasia"

MITRE

CAT: Specialized Matching for Entities

- **For term search and highlighting in text, entity-specific search strategies may improve retrieval results by accommodating**
  - Mixture of translation and transliteration
    - E.g. looser match criteria for transliterated elements vs. "real words"
  - Entity specific stopwords
  - Abbreviation-to-full form matching

Fuzzy matching in the MemoQ CAT tool



Retrieved from http://www.translationtribulations.com/2013/06/understanding-fuzzy-term-matching-in.html 09/03/2014

CAT: Inflection

- **For inserting known terminology translations into text, CAT tools *may***
  - Detect inflected forms of terms
  - Allow translators to insert translations with appropriate inflections

Retrieved from http://www.udel.edu/fllt/instruction/atajoch1.html 09/03/2014

Entities in Context

Strategies for Translating Entity Names in Context

- **No special handling: just get enough data**
  - Google's scores on transliterations of low frequency names illustrate the limitations of this approach (You et al. 2012)
  - Microsoft researchers claim that no special handling they have tried improves entity translation more than increasing the quantity of training data
- **Basic approaches**
  - Entity names identified for special handling when text is processed by MT system vs.
  - Entity name translation is integrated with the rules or statistical models of the MT system
  - Reliance on bilingual lexicons vs. learning

MITRE

Finding Entity Names in Context

- **Special handling for entity names requires procedures to recognize them in the source input**
- **Challenges of entity extraction are well known**
- **Errors cascade from inaccurate extraction results**
  - Appropriate handling of entity names requires accurate recognition and classification of entity type (personal name, location, organization, etc.)
  - An experienced MT researcher has stated that extraction must achieve 92-93% accuracy in order for special handling of entity names to improve MT and lower accuracies can be detrimental
- **After recognizing and classifying, it is still necessary to decide whether the entity name (or parts of the name) should be transliterated**

Special Handling Example: 2012 Raytheon BBN Patent (Weischedel 2008)

- **An entity extraction system extracts the entity names and their types, leaving placeholders in the source text**
- **Entity names are processed according to their types**
  - Rules for dates and times
  - Transliteration for person names
  - A mixture model that uses bilingual dictionary resources to assign a probability to the name translation using a tunable weight associated with the dictionary
- **The text with placeholders is translated using a phrase-based SMT model**
  - The probabilities associated with the entity names are merged with the probabilities assigned by the SMT model to the sentence
  - An incremental process finds the most probable translation using constraints to ensure that the words in entity phrases are kept together

Another Special Handling Example

- **Okuma et al. (2007) substitute source names not in the phrase table with high frequency source names of the same type**
  - Translation proceeds as usual
  - Then they replace the high frequency names with translations of the source names from a bilingual lexicon
- **Achieved significant improvements in BLEU scores for test sets with high frequencies of names**
  - Japanese to English translations of sentences with location and person names improved more than 4 BLEU points for location names and more than 3 BLEU points for person names
  - English to Japanese translations improved almost 4 BLEU points for person names but decreased slightly for locations
- **Using placeholders in both examples preserves the context for translation of the surrounding text**

MITRE

Special Handling without Extraction

- **Hermjakob et al. (2008) train a classifier to recognize words that should be transliterated**
  – Eliminates need for named entity recognizer
  – Addresses the problem of deciding, once a name is recognized, whether it should be transliterated
  – Achieved F score of 0.94 on a test set
- **During training, names which have been tagged as words that should be transliterated are transliterated**
  – The transliterations are added to the phrase table with a special feature set to a value of 1
  – The value is adjusted along with other feature weights in the tuning process
- **90% of entity names in an Arabic text were correctly translated into English**

MITRE

A Simple Approach: Add Names

- **Add bilingual name lexicons to the training data**
  - This is a variant of the “get more data” strategy
  - Instead of special handling, add special data
- **Pal et al. (2010) improved English to Bangla translations almost 5 BLEU points for travel texts**
  - Automatically aligned entity names in the training data using a transliteration similarity score
  - Added the aligned names to the training data
- **Large improvements in BLEU are not typical**
  - Both Okuma et al. and Pal et al. used test data with many entity names
  - Pal et al. used a relatively small training set so that adding the aligned names significantly increased the size of the training set

General OOV Approaches

- **Pal et al. (2010) experimented with concatenating all of the name parts into a single “word”**
  – This is a general strategy for mapping multi-word source expressions to multi-word target expressions
  – No significant BLEU score increase
- **Transliteration is one of 4 procedures Habash (2008, 2009) uses to handle expressions that are not in the phrase table (OOV)**
  – Possible transliterations are added to the phrase table with low translation probabilities
  – All 4 procedures are applied to all OOV expressions
  – Transliteration alone increased BLEU score 0.4 points
  – All 4 procedures increased BLEU score 1.4 points

Summary of Recent Approaches

Researchers	Description	Translates names in context	Transliteration on the fly vs. add dictionary	Improvement in BLEU scores
Raytheon BBN patent (2012)	Translate names separately with placeholders in context	yes	yes	n/a
Pal et al. (2010)	Add names to training set	no	no	+4.6
Habash (2009)	Transliterate unrecognized expressions, add to phrase table with low probabilities	no	yes	+1
Hermjakob et al. (2008)	Recognize names to transliterate, add to phrase table with a feature	yes	yes	n/a
Okuma et al. (2007)	Substitute name with more frequent name (same type) for translation, then replace	yes	no	+0 - +4.2

Evaluation of Entity Translation

MITRE

What Makes a Good Evaluation?

- **Objective – gives unbiased results**
- **Replicable – gives same results for same inputs**
- **Diagnostic – can give information about system improvement**
- **Cost-efficient – does not require extensive resources to repeat**
- **Understandable – results are meaningful in some way to appropriate people**

Framework for Evaluation: EAGLES 7-Step Recipe/ISLE ➔ FEMTI

1. **Define purpose of evaluation – why doing the evaluation**
2. **Elaborate a task model – what tasks are to be performed with the data**
3. **Define top-level quality characteristics**
4. **Produce detailed system requirements**
5. **Define metrics to measure requirements**
6. **Define technique to measure metrics**
7. **Carry out and interpret evaluation**

http://www.issco.unige.ch:8080/cocoon/femti/st-home.html

Evaluation in Context

Both Component-level and System-level Evaluation are necessary

- Evaluation dependent on use case
- Is the desired result:
  - **CLIR**: The ability to retrieve the set of all (unstructured) document holdings containing a mention of an individual
  - **Structured Data Retrieval / Management**: The ability to retrieve the set of transliterated or translated name records, linked to information about individuals, organizations or locations
  - **Link analysis**: The ability to visualize the set of relationships between (resolved) identities / entities in potentially multilingual organizational holdings
  - **Triage**: The ability to have humans identify whether people, organizations, or locations of interest are mentioned in a document, and what role they play.
- Use case and evaluation are related but different for each of the above
  - Each has translation or transliteration component to evaluate as well as the end-to-end system evaluation (which may contain identity matching/resolution and other information retrieval components).

Evaluation for Named Entities in MT

- **BLEU and other completely automated metrics don't accord special importance to named entities**
  - Systems have improved BLEU scores by deleting NEs or NFWs from output
- **IR-based use cases for both structured and unstructured information**
  - Based on TREC (IR) Methodology
  - Results pooling with human annotation based on guidelines
  - Precision, Recall, F-measure
  - Other metrics possible
- **Miller and Vanni recommend specific evaluation of Named Entity Translation (PLATO – Predictive Linguistic Analysis of Machine Translation Output)**
- **Link Analysis or Knowledge-Base Population may benefit from metrics for clustering evaluation**
  - **NIST TAC KBP Track on Entity Linking** 2014:
    - (http://nlp.cs.rpi.edu/kbp/2014/)
  - **NIST TAC KBP Track on Slot Filling**: 2014:
    - (http://surdeanu.info/kbp2014/def.php)

MITRE

Basic Metrics: Precision and Recall

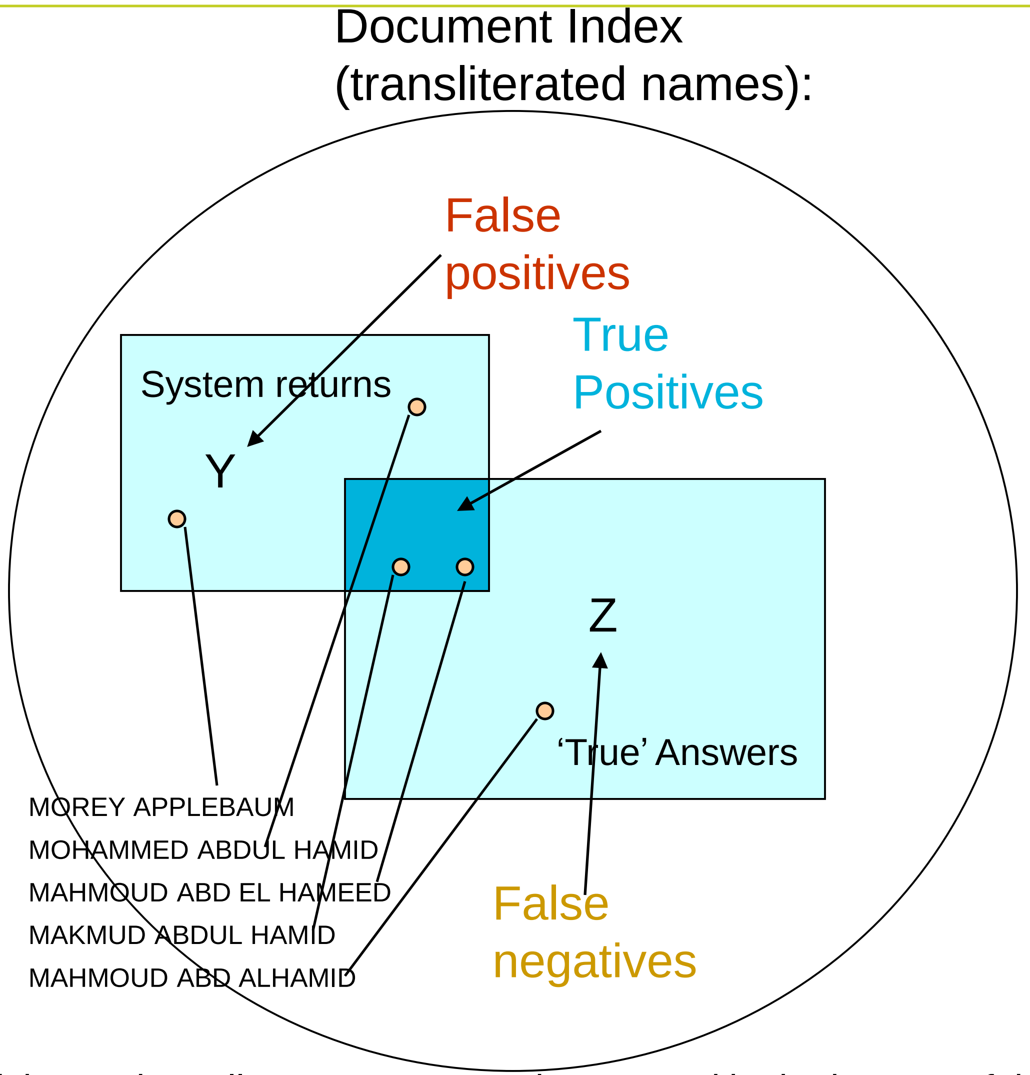
Query:

MAHMOUD ABDUL HAMEED

12/10/1945

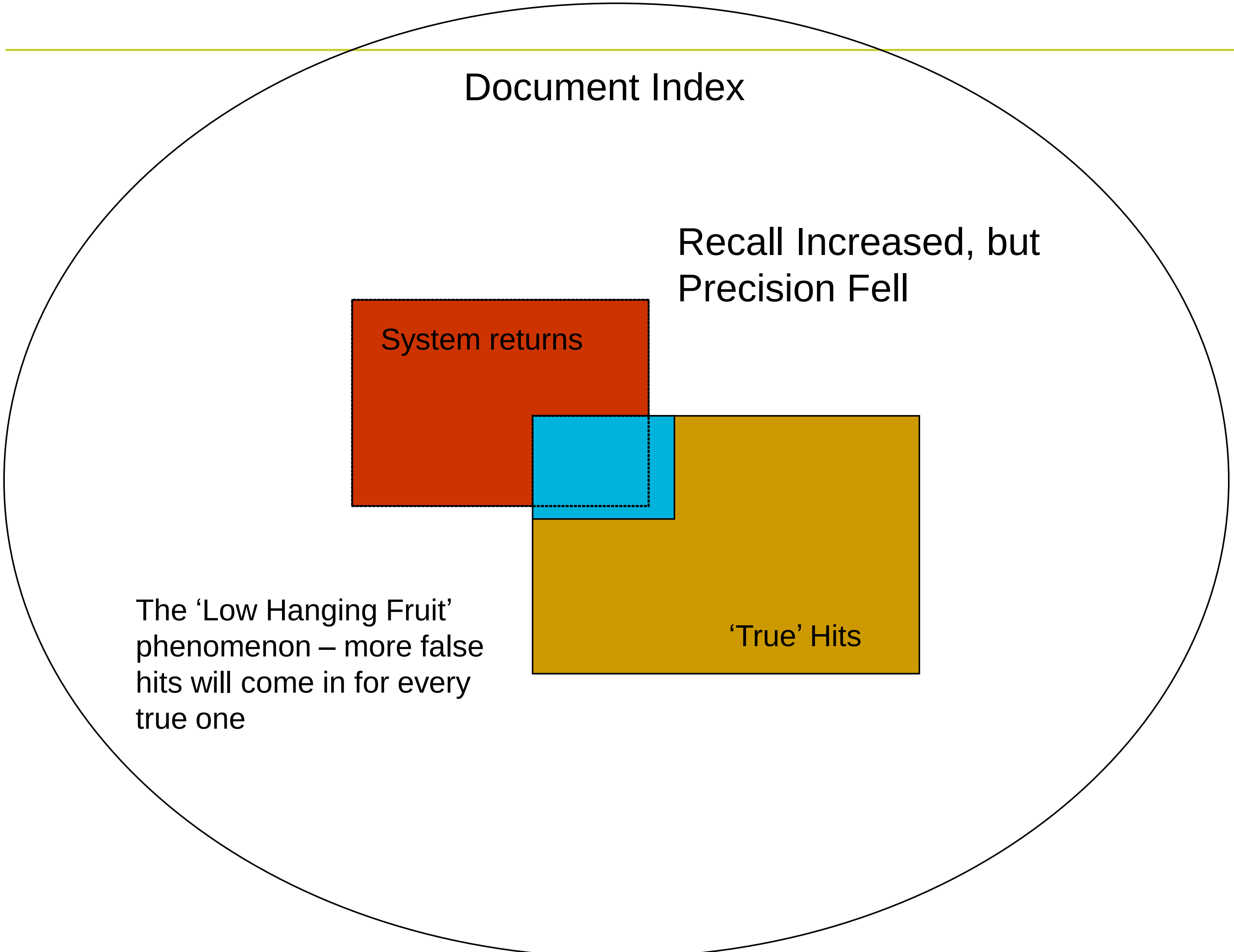
Precision (P) = X/Y (2/4)

Recall (R) = X/Z (2/3)



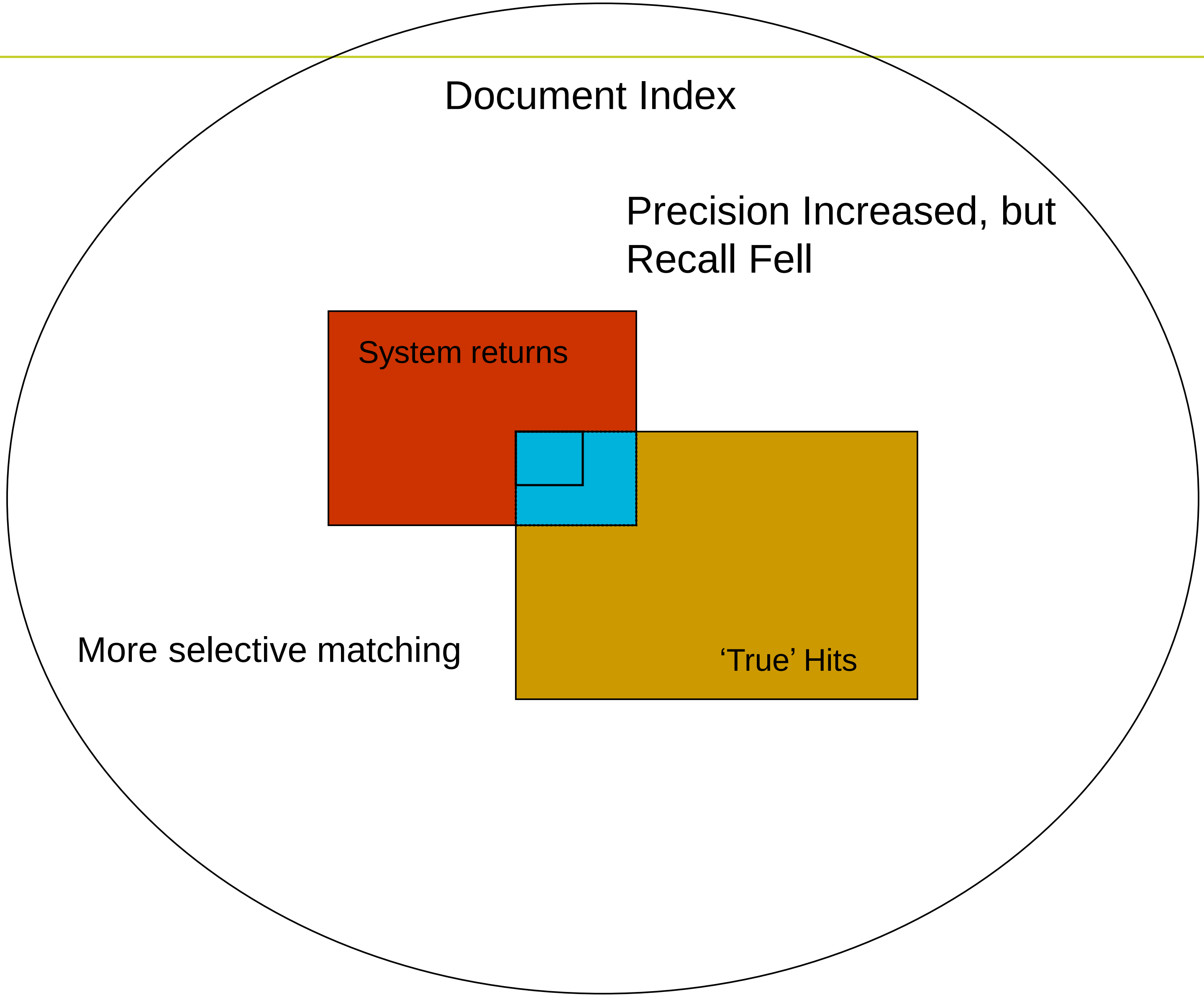
Note: Other metrics are possible; precision and recall are common, and presented in the interest of time.

Precision and Recall Inversely Related (1)



MITRE

Precision and Recall Inversely Related (2)



MITRE

Sample Evaluation Metric: F-score combines Precision and Recall

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}$$

- 3 values of Beta:
  - F_1 – Standard, Balanced F-Score = 2PR / P + R
  - F_2 – Favors Recall
  - $F_{0.5}$ – Favors Precision

MITRE

Another Possible Metric: MAP

- Mean Average Precision: Unlike F-score, rank order of results counts
  - All queries contribute equally
  - Unreturned matches count against you
  - Scores can be anything (tie-friendly algorithm)
  - Diminishing returns for low-level matches

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP(q)}}{Q} \qquad \mathrm{AveP} = \frac{\sum_{r=1}^{N} (P(r) \times \mathrm{rel}(r))}{\text{number of relevant documents}}$$

$$\mathrm{P(r)} = \frac{|\{\text{relevant retrieved documents of rank r or less}\}|}{r}$$

MITRE

Exercises on Contextualized Evaluation of MT of Named Entities

- **(handout of example translations)**

References: Entity Name Translation (1)

Al-Onaizan, Y., & Knight, K. (2009, August 25). Patent No. 7,580,830. US. Patent and Trademark office. Retrieved from http://www.google.com/patents/US7580830.

Babych, B., & Hartley, A. (2003). Improving machine translation quality with automatic named entity recognition. 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT, (pp. 1-8).

Feng, D., Lü, Y., & Zhou, M. (2004). A new approach for English-Chinese named entity alignment. Proceedings of the Empirical Methods in Natural Language Processing, (pp. 372-379).

Habash, N. (2008). Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, (pp. 57-60).

Habash, N. (2009). REMOOV: A tool for online handling of out-of-vocabulary words in machine translation. Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR).

Hermjakob, U., Knight, K., & Daumé III, H. (2008). Name translation in statistical machine translation-Learning when to transliterate. Proceedings of the Association for Computational Linguistics, (pp. 389-397).

Huang, F., Vogel, S., & Waibel, A. (2004). Improving named entity translation combining phonetic and semantic similarities. Proceedings of Human Language Technology-North American Association for Computational Linguistics, (pp. 281-288).

Ji, H. (2009). Mining name translations from comparable corpora by creating bilingual information networks. Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora, (pp. 34-37).

References: Entity Name Translation (2)

Jiang, L., Zhou, M., Chien, L. F., & Niu, C. (2007). Named entity translation with web mining and transliteration. Proceedings of the International Joint Conference on Artificial Intelligence, 7, (pp. 1629-1634).

Lam, W., Chan, S.-K., & Huang, R. (2007). Named entity translation matching and learning: With application for mining unseen translations. ACM Transactions on Information Systems, 25(1), Article 2.

Lin, W.-P., Snover, M., & Ji, H. (2011). Unsupervised language-independent name translation mining from Wikipedia infoboxes. Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), (pp. 43-52).

Okuma, H., Yamamoto, H., & Sumita, E. (2007). Introducing translation dictionary into phrase-based SMT. Proceedings of the MT Summit XI, (pp. 361-368).

Pal, S., Kumar Naskar, S., Pecina, P., Bandyopadhyay, S., & A., W. (2010). Handling named entities and compound verbs in phrase-based statistical machine translation. Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications

Sproat, R., Tao, T., & Zhai, C. (2006). Named entity transliteration with comparable corpora. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, (pp. 73-80).

You, G. W., Hwang, S. W., Song, Y. I., Jiang, L., & Nie, Z. (2012). Efficient entity translation mining: A parallelized graph alignment approach. ACM Transactions on Information Systems (TOIS), 30(4), Article 25.

Weischedel, R., Xu, J., & Kayser, M. (2008). Patent No. 20080215309. BBN Technologies Corp. Cambridge, MA, US. Retrieved from http://www.freepatentsonline.com/y2008/0215309.html.

References: Transliteration (1)

AbdulJaleel, N. and Larkey, L. 2003. Statistical transliteration for English-Arabic cross language information retrieval. In Proceedings of the Conference on Information and Knowledge Management. New Orleans, LA, pp. 139-146.

Al-Onaizan, Y. and Knight, K. 2002a. Machine translation of names in Arabic text. In Proceedings of the ACL Conference Workshop on Computational Approaches to Semitic Languages.

Al-Onaizan, Y., & Knight, K. 2002b. Translating named entities using monolingual and bilingual resources. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (pp. 400-408). Association for Computational Linguistics.

Gao, W., Wong, K., and Lam, W. 2004. Phoneme-based transliteration of foreign names for OOV problem. In Proceedings of First International Joint Conference on Natural Language Processing.

Goto, I., Kato, N., Uratani, N., & Ehara, T. 2003. Transliteration considering context information based on the maximum entropy method. In Proceedings of MT-Summit IX.

Jaro, M. A. 1995. Probabilistic linkage of large public health data files (disc: P687-689). Statistics in Medicine 14:491–498 (matching reference)

Jung, S. Hong, S., and Paek, E. 2000. An English to Korean transliteration model of extended Markov window. In Proceedings of COLING.

Kang, B. J., & Choi, K. S. 2000. Automatic transliteration and back-transliteration by decision tree learning. In Proceedings of the 2nd International Conference on Language Resources and Evaluation.

MITRE

References: Transliteration (2)

Knight, K. and Graehl, J., 1997. Machine Transliteration, In Proceedings of the Conference of the Association for Computation Linguistics (ACL).

Li, H., Zhang, M., & Su, J. 2004. A joint source-channel model for machine transliteration. In Proceedings of Conference of the Association for Computation Linguistics (ACL).

Meng, H., Lo, W., Chen B., and Tang, T. 2001. Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval. In Proceedings of ASRU.

Oh, J. H., & Choi, K. S. (2002, August). An English-Korean transliteration model using pronunciation and contextual rules. In Proceedings of the 19th international conference on Computational linguistics-Volume 1 (pp. 1-7). Association for Computational Linguistics.

Oh, J. H., & Choi, K. S. (2006). An ensemble of transliteration models for information retrieval. Information Processing and Management, 42, 4, 980-1002.

Virga, P. and Khudanpur, S. 2003. Transliteration of proper names in cross-lingual information retrieval. In Proceedings of the ACL Workshop on Multi-lingual Named Entity Recognition.

Wan, S. and Verspoor, C. 1998. Automatic English-Cinese name transliteration for development of multilingual resources. In Proceedings of the Joint Meeting of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics.

Winkler, W. 2002. Record linkage and Bayesian networks. In Proceedings of the Section on Survey Research Methods, American Statistical Association. Retrieved as RRS2002/05 from http://www.census.gov/srd/www/byyear.html. (matching reference)

MITRE