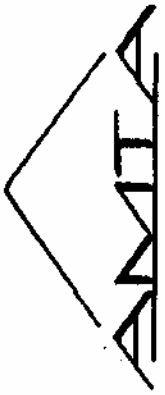


Association for Machine Translation in the Americas

AMTA - 2006 CONFERENCE

TUTORIAL ON

NAME TRANSLATION

Presenters:
Keith Miller and Sherri Condon
The MITRE Corporation

BOSTON MARRIOTT CAMBRIDGE
CAMBRIDGE, MA

8 - 12 AUGUST 2006

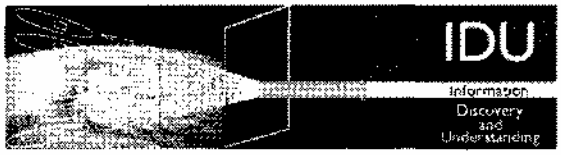
Name Transliteration

Current Methods, Applications, and Evaluation in Name Transliteration and Translation

Sherri Condon
Keith J. Miller
The MITRE Corporation

August 8, 2006

Washington C3 Center



MITRE

First Activity: Morning Calisthenics

Overview

1. Why focus on names and name transliteration? What are problems?
   - Why are names such a challenge?
     - across languages, scripts, and cultures
   - Survey of problems with a focus on Arabic names

   Transliteration issues:
   - Transliteration vs. character mapping
   - Competing transliteration schemes and standards
   - Methods for automatic transliteration

   Survey of matching approaches
   - advantages / disadvantages of each
     - Matching across scripts
     - Methods for data acquisition
     - Transliteration
     - Phonological interlingua

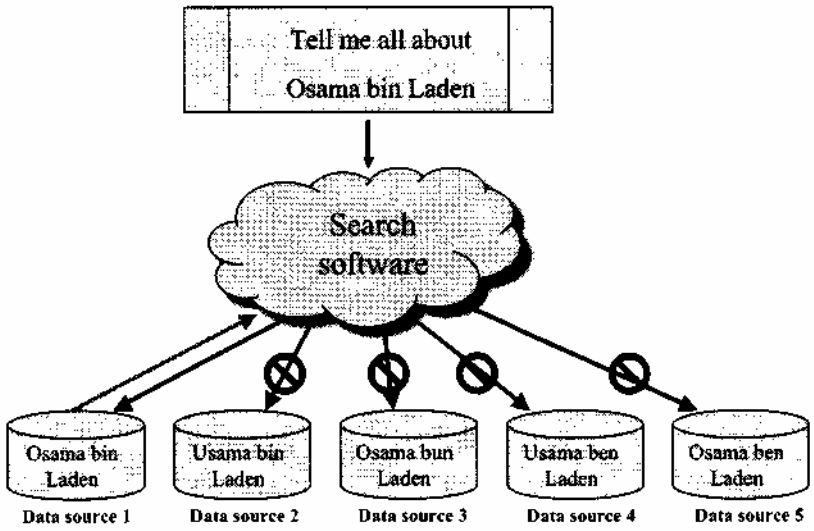
   Case study comparing matching systems for Romanized Arabic names

Overview

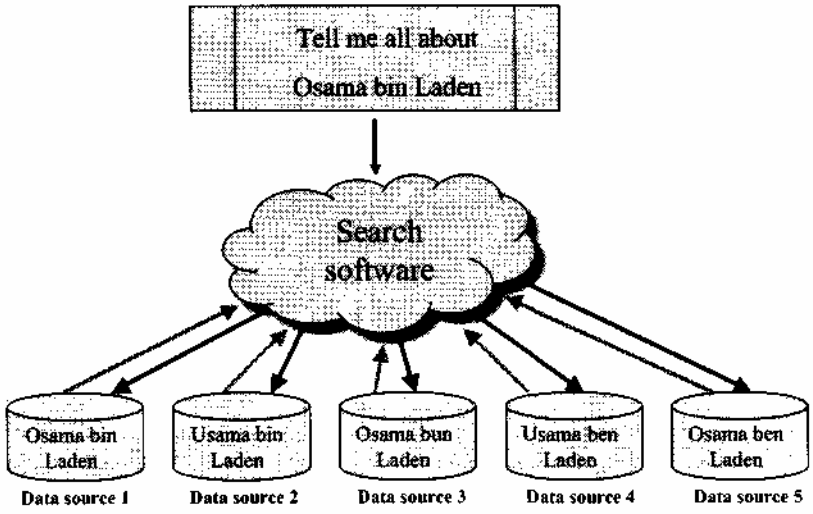
2. Possible Solutions and Evaluation

- Entity Extraction to Identify Names for Transliteration
- Evaluation of Name Search and Matching Systems
  - Development of ground-truth sets
    - Human adjudication
    - Estimation techniques
- Case study / Activity: adjudication exercises
- Metrics and Inter-adjudicator agreement
- Evaluation metrics for names in MT
  - holistic vs discrete point evaluation
  - Activity: Evaluation Exercise / names in MT
- Performance and other considerations

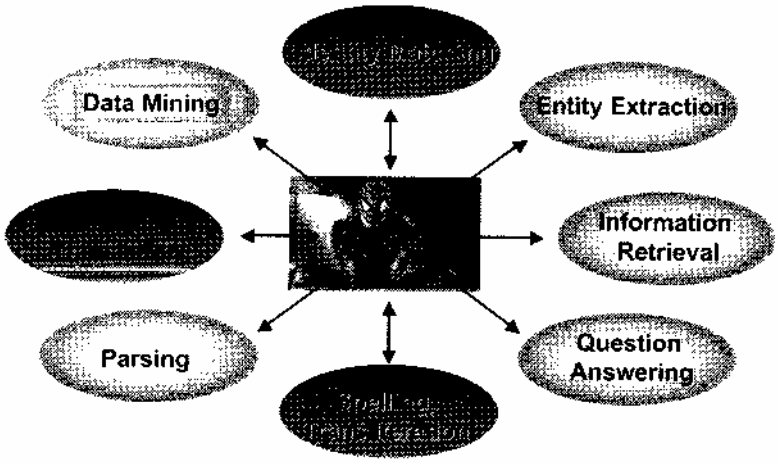
Without Proper Name Handling



With Proper Name Handling



Impact of Proper Name Processing



Problem Sources

Data Acquisition

- Spoken sources (a quiz)
- Variation in written sources, e.g. Open Source

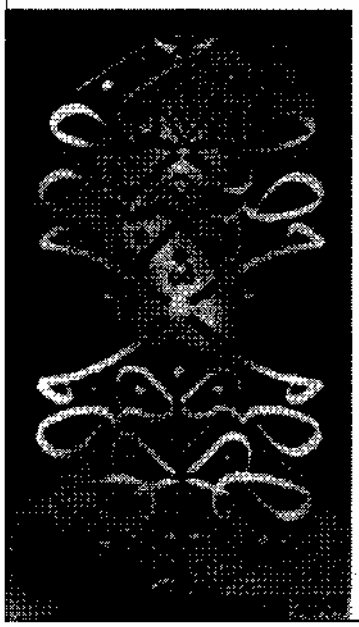
Data Exchange / Data Quality

- Differing data models between systems
- Ill-defined or non-existent standards for data exchange

Differing Cultural and Linguistic Conventions Regarding Names

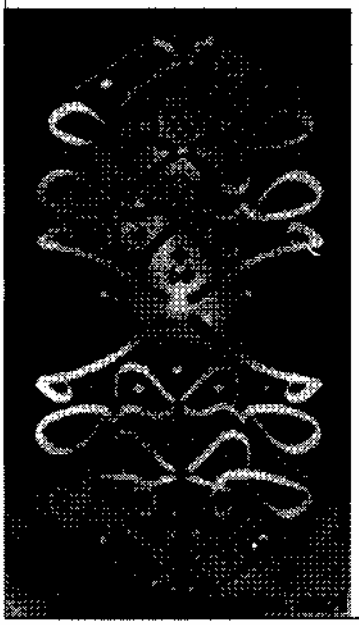
- Syntax
- Morphology
- Non-Roman Scripts and Transliteration

Complications for Translation



- Translation / Transliteration Decision
  - Proper Noun / Common Noun class overlap
  - Borrowing
- Transliteration standard(s) / method(s)
  - Readability
  - Reversibility

Complications for Consistency in Representation



- Initials
- Nicknames *(Bob, Pat, Patti)*
- Name Variants
- Titles *(COL, Dr.)*
- Qualifiers *(Jr., II)*
- Particles *(von, de, bin, abu)*
- Prefixes *(Mc/Mac, al)*
- Suffixes *(-vich, -ovic, -ov)*
- Absent Name Parts
- Incorrect Fielding
- Name Structure

Phonebook from Montgomery County, MD

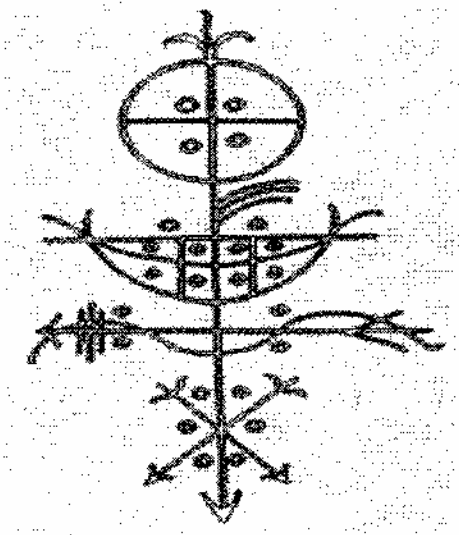
Czajka

Dahlke

Encinas

Mahdjoor

Ndongku



Sforza

Sfrisi

Sgaggero

Boudreaux

Thibodaux

Second Activity

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

Arabic Name Structure (Simplified)

- 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"

Transliteration: Another Dimension of Variation

- **Multiple transliteration standards & traditions:**
  Francophone traditions (Wasim = Ouassime), Legacy data
- **Acoustic errors: Ali = 'Ali**
- **Dialectal variants: Bourguiba = Abu Ruqayba**
- **Non-native names: Pavel = Bafil**
- **Segmentation: 'Abd Al Rahman ='Abdurrahman**
- **N-to-n mappings: Walid = ولید and والد**
- **Missing Information: محمد = "m" "h" "m" "d"**

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; Mohammed
- Transliteration / Romanization from different scripts introduces other challenges
- May have completely different naming model
  - no concept of surname
- Complication for ID matching in general:
  - Lack of emphasis on record keeping: e.g. inexact or unavailable birth dates

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 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

Proper Names are Special

- 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
- In many languages, names have complex structures
- Models that work for the rest of the language may not be effective for names

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
- In many contexts these two types are incomplete because additional languages are involved, e.g. transliterating a Chinese name from Arabic into English

Transliteration Standards

- Most transliterations are *de facto*
- Many competing standards
  - Government: FBIS, SATTS, IC, BGN (place names)
  - Academia, industry
- Problems of adoption/enforcement of standards
  - Readability vs. reversibility
  - Limitations on character sets
  - Linguistic controversies (e.g., phonemic vs. phonetic)
  - Legacy data
- Multiple encoding standards, too

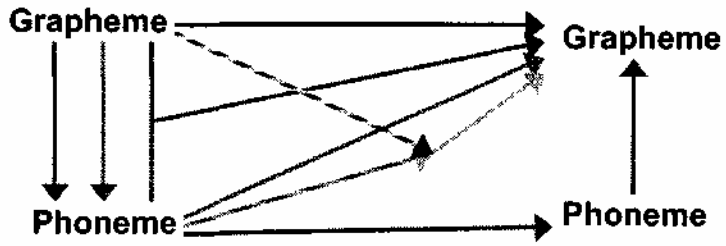
Arabic Transliteration Standards from Wikipedia

- Deutsche Morgenländische Gesellschaft (1936)
- ISO/R 233 (1961). Replaced by ISO 233 in 1984 but still encountered.
- BS 4280 (1968): Developed by the British Standards Institute. [2]
- SATTS (1970s)
- UNGEGN (1972)
- DIN-31635 (1982)
- ISO 233 (1984)
- Qalam (1985)
- ISO 233-2(1993)
- Buckwalter Transliteration (1990s)
- ALA-LC (1997)
- SAS: Spanish Arabists School

Automatic Transliteration

- Recent attention has led to a variety of approaches for major languages with non-Roman scripts: Chinese, Korean, Japanese, Arabic
- Most methods use probabilities acquired from data sets
- Evaluating methods is not straightforward
- Human transliteration can be far from perfect

Automatic Transliteration Choices



1. Grapheme to grapheme
2. 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

- **Advantages**
  - Direct text input and output
  - Captures influence of graphemic representation on transliteration, e.g. Graham as غراهام = ġrāhām instead of as غرام = ġrām (Onaizan & Knight, 2002a)
- **Disadvantages**
  - Transliteration often reflects pronunciation, e.g. Knight as نايت = nāyt (Onaizan & Knight, 2002b)
  - Spelling may be irregular, especially in English

Grapheme to Grapheme Example: Onaizan & Knight (2002b)

- 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)
  - Extension includes letter trigram model for words not on lists
- Transliteration maximizes $P_s(w|a) \sim P(w)\ P(a|w)$
- *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

Grapheme → Phoneme → Grapheme

- Advantages
  - Avoids spelling irregularities (for source language)
  - Captures sound correspondences (for source language)
- Disadvantages
  - Requires mapping graphemes to phonemes in source language, which is subject to spelling irregularities
  - Advantages apply only to source language

Grapheme → Phoneme → Grapheme Example: Onaizan & Knight (2002b)

- 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)*

Grapheme & Phoneme to Grapheme

- Advantages
  - Captures both sound and grapheme information (for source language)
  - Addresses spelling irregularities
- Disadvantages
  - Requires mapping graphemes to phonemes in source language
  - Advantages apply only to source language

Grapheme + Phoneme: Two Ways

- Grapheme – phoneme correspondence in L1 maps to grapheme in L2
  - L1 grapheme/phoneme association is analogous to adding context information
  - Example: Oh & Choi (2002, 2005) English to Korean
- Grapheme – grapheme and phoneme – grapheme probabilities are combined
  - Example Onaizan & Knight (2002b)
  - $P(w|a) = \lambda P_s(w|a) + (1 - \lambda) P_p(w|a)$

Grapheme → Phoneme → Phoneme → Grapheme

- Advantages
  - Captures sound correspondences for both languages
  - Avoids spelling irregularities for both languages
- Disadvantages
  - Does not capture orthographic correspondences
  - Requires mappings between phonemes and graphemes, which are subject to spelling irregularities

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

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

Variations

- Handcrafted mappings
  - 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 (2005) 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 representations	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

Transliteration Enhancements

- Onaizan & Knight (2002b) extend the candidate list by searching for name parts, e.g. * Annan
- Web Frequencies to rank candidates
  - Oh & Choi (2005) and Onaizan & Knight (2002b) use normalized Web counts to rescore candidates
  - Onaizan & Knight (2002b) also use contextual web counts: name plus title or key words or local terms
  - Oh & Choi (2005) search for source/transliteration pairs as phrases or in the same document (for chemical names)
- Onaizan & Knight (2002b) identify coreferent sub-phrases and rank according to the longer referents

Alternative Web-Based Approach

- 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
  - Language independent scoring based on common sound change and second language learner errors
  - Implemented as costs of substitution/deletion/insertion
- Also score candidates based on frequency profile
- Combine similarity and frequency scores

Computation of Frequency Profile

- Sproat, Tao & Zhai (2006); Tao et al. (2006)
- Pool all news documents from a single day into one pseudo-document and compute the frequency of each candidate (both languages) in each day
- Normalize the raw frequency vector to a distribution over all of the time points (days)
- Use Pearson correlation coefficient to compute similarity of vector distributions
- Experimenting with score propagation to increase weights when higher numbers of candidate pairs co-occur in the same document

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
  - غورال al qur al gur

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 and Precision
- 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 is 29.9%, C to E is 62.1%

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

Overview of Matching Strategies

- Search Keys
- String Similarity
- Token Based
- Variant Look-up
- Variant Generation
- Normalization
- Intelligent Search and Match

Search Keys: Soundex

Code	Characters
0	a e h i o u w y
1	f p v
2	c g j k q s x z
3	d t
4	l
5	m n
6	r

Examples:

Rodriguez → R362
Li → L
Lee → L
Lu → L

1. Replace all but the first letter of s by its phonetic code.
2. Eliminate any consecutive repetitions of codes.
3. Eliminate all occurrences of code 0 (that is, eliminate vowels).
4. Return the first four characters of the resulting string.

Soundex Problems

- Variantions: Phonix - 160 transformations - (Gadd, 1990), Phonex (Lait & Randell, 1996), Editex (Zobel & Dart, 1996), Ipadist (Zobel & Dart, 1996), Metaphone (Philips, 1990), NYSIIS (NY State Identification and Intelligence System)
- Dependency on initial grapheme or sound
- Different names have same code (collisions)

  Mohammad → M530 Mahmoud → M530
- Variant spellings of a single name have different codes (false negatives)

  Abdel Rahman → A134 or R550

  Abdurrahman → A136
- Lait and Randell (1996) found that Soundex and Metaphone identified only 30-40% of true matches in published name data and less than 20% of true matches in deliberately corrupted samples

String Similarity: N-Grams

- **Ukkonen (1992)** distance (s,t) = $\sum_{g \in G_s \cup G_t} |\, s[g] - t[g] \,|$

  G_x = set of n-grams in string x,

  x[g] = number of occurrences of n-gram x in g

- **Approximation (Zobel & Dart, 1995)** = $|\, G_s \,| + |\, G_t \,| - 2\,|\, G_s \cap G_t \,|$
- **DICE coefficient** = $2\,|\, G_s \cap G_t \,| \,/\, (\,|\, G_s \,| + |\, G_t \,|\,)$, G_x = set of bigrams in string x
- **Jaro**

$$Jaro(s,t) = \frac{1}{3} \cdot \left(\frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - T_{s',t'}}{|s'|} \right)$$

  s′ and t′ are the characters in common (in order, with limit on position distance)

  $T_{s',t'}$ is half the number of positions without characters in common

- **Jaro-Winkler**

$$Jaro\text{-}Winkler(s,t) = Jaro(s,t) + \frac{P'}{10} \cdot (1 - Jaro(s,t))$$

  P is the length of the longest common prefix of s and t, P′ = max(P,4)

Properties of N-Gram Distance

- **Best if include begin and end tokens**
- **Useful for retrieval**
  - $|\, G_s \cap G_t \,|$ can be computed from an inverted index
  - High performance coarse matching strategy
- **Compares to edit distance for some corpora**
- **Not effective on short strings**

String Similarity: Edit Distance

- Classic Levenshtein
  - Insertion
  - Deletion
  - Substitution
- Relatives of Levenstein
  - Needleman-Wunsch (Needleman & Wunsch, 1970)
  - Monge-Elkan (Monge and Elkan, 1996, 1997)
  - Smith-Waterman (Smith and Waterman, 1981)
- Use dynamic programming for alignment
- Developed for comparison of biosquences

Example: Smith
Smythe
1 substitution, 1 insertion
Edit distance = 2

Properties of Edit Distance

- High processing costs
- Effective on many corpora
- Not effective on short strings
- Distance penalties can be learned (Bilenko and Mooney, 2003)
- Must be normalized to be useful

Token-Based Measures

- Jaccard
- TF.IDF
- Jensen-Shannon (weights n-grams as smoothed probability of the token appearing in the corpus)
- Fellegi-Sunter (1969)

  odds ratio log(Pr(Match|*s*,*t*) / Pr(Notmatch|*s*,*t*)
- Hybrid token-string
  - SoftTFIDF (Cohen, Ravikumar, & Feinburg, 2005)
  - Monge-Elkan (1996,7) recursive matching scheme

Variant Look-up

- Dictionaries, gazetteers
- Useful for dissimilar variants
  - *Elizabeth* and *Betsy*
  - *Richard* and *Dick*
  - *Mohammad* and *Mhd*
- Required for some linguistic knowledge
  - Titles, prefixes, suffixes
  - Particles, qualifiers
- Limited to contents of look-up tables
- Requires space to store, time to search

Variant Generation

- From look-up tables or rule-based
- Can be used to produce search keys
  - Keys generate multiple indices into database
  - Effective coarse matching for first cut
- Problems
  - Exact match of variants against names is brittle
  - Numbers of variants can become enormous
    - Using COTS Arabic name variant generator and attested Arabic names, the sum of variants for each name part of the full names ranged from 32 to 213,825,733
    - The number of possible combinations of variant name parts for the full names ranged from 281,344 to 1,879,956,196,216
  - Limited to variants anticipated by rules and tables

Normalization: Methods

- Rule-based: pattern matching transforms
- Normalized form based on
  - Morphological analysis
  - Predictability ("deep" structure)
  - Salience (consonants vs. vowels)
  - Phonetic or graphemic representation
  - Emics of cultures involved
- Need not be human readable
- Can be keys or search arguments
- Problems
  - Rule interaction and ordering grows out of control
  - Highly knowledge intensive

Intelligent Search and Match

Incorporate linguistic knowledge to adjust settings for cultures and purposes

- Strategies for name structures
  - Wheeling
  - Missing/additional parts
  - Initials and other special handling
- Tuning parameters of similarity metrics
  - Weights (name parts, gender differences)
  - Thresholds
    - Depth of search, exhaustiveness of search
    - Number of returns
    - Cuts and coarse matching strategies
- Exploitation of reference tables
- Generation of keys, selection of algorithms

Name Matching "Cocktails"

- Combine results from several methods
  - Classifiers
  - Adaptive Matching and clustering (Cohen & Richman, 2001; 2002)
  - Classical record linkage (Fellegi & Sunter, 1969; Winkler, 2002)
  - MARLIN: Multiply Adaptive Record Linkage with Induction (Bilenko & Mooney, 2003)
  - Generative probability models (Pasula et al., 2002)
- Require appropriate training data
- Combining metrics and additional information for identity matching is a promising research area

Matching Across Scripts

Variants work both ways: Vojislav Kostunica

كوشتونيتسافويسلاف fwyslAf kwshtwnytsA

كوستنيتشفويسلاف fwyslAf kwstnytsh

كوستنيتشافويسلاف fwyslAf kwstnytshA

كوستونيكافويسلاف fwyslAf kwstwnykA

كوستونيتشفويسلاف fwyslAf kwstwnytsh

كوستونيتشافويسلاف fwyslAf kwstwnytshA

كوستنيستافويوسلاف fwywslAf kwstnystA

Cross Script Matching: Data Needs

- Ideal: name n-tuples

  艾伦

  Alan 阿兰 علان

  阿朗

- Another problem: negative examples for learning
- Transliterate and search
- Entity taggers with parallel corpora
  - Bootstrap on excellent English taggers
  - Minimal: candidate lists based on segments
  - Maximal: MT alignment methods
  - Incremental: preliminary matching

Exploit Parallel Corpora with Matching

L1W1 L1W2 L1W3 L1W4 L1W5 L1W6 L1W7 L1W8

L2N1 L2N2

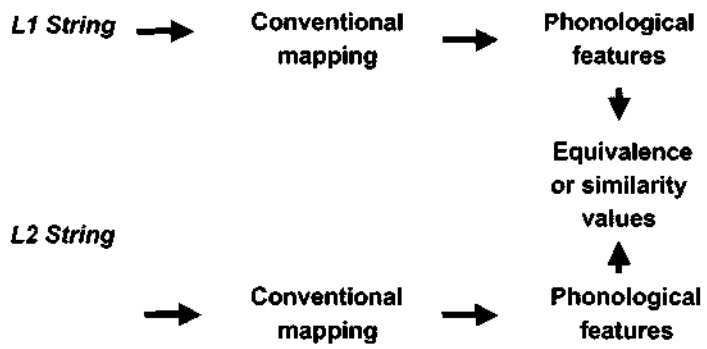
Select the match with the highest similarity score in the text unit (sentence, paragraph)

Transliteration

- Names must be in the same character set for similarity measures so character mapping is necessary for string comparison
- Automatic transliteration is not suitable for matching
  - At best another variant
  - At worst adds unpredictable noise
- Instead of transliterating, a matcher can assign probabilities to potential matches
- Probably best as input to a classifier
- Most automatic transliteration methods require training data

Phonological Interlingua

Exploit grapheme-phoneme correspondences available in traditional grammars



Phonological Interlingua Considerations

- Does not require training corpora
- Can employ existing similarity metrics
  - String similarity or specialized for IPA (Ipadist)
  - ALINE (Kondrak, 2000)
- Can be based on native phonemes or reflect non-native perceptions
- Can incorporate learning methods and be mixed into cocktails
- Problem: not all transliteration is based on sound correspondences. There are semantic and graphemic correspondences, too.

More Empirical Questions

- Very short names (Asian)
- Very long names with repeat n-grams (Thai, Hawaiian)
- Role and value of various features
  - Gender
  - Ethnicity: of name vs. script
- General methods to exploit name structures

MITRE Evaluation: Matching Romanized Arabic Names

- Test set of 660 queries matched against 8,792 names
- 2 native speakers of Arabic generated 2,800 variant database names (each associated with query name)
- Created several variants of query name

  Muhammad Abd al Amir

  → Mohamed Abdel Amer, Mhammad Abdul Amir
- Created one "false positive" for each

  (e.g., Mahmud Abdul Amir)
- For each query, database contained at least 3 pre-judged matching names and 1 similar but false match for initial ground truth

  independent of pooled matching results

Matching Eval: More Methodology

- All systems set parameters for Arabic names, though not necessarily specialized
- No analysis of database by vendors was permitted, so not "optimized" but allowed multiple runs at different settings
- Vendors provided preferred settings (data flow, system architecture, etc.)
- All returns were manually judged by the 2 Arab consultants
- Inter-judge agreement rate: 86%
- Judgment (truth) database was updated for each iteration, with fewer new judgments required for each iteration
- Metrics: precision, recall, F score at top 5 and top 10 returns

MITRE Matching Evaluation: Scoring

- All returns were manually judged by the 2 Arab consultants
- Inter-judge agreement rate: 86%
- Judgment (truth) database was updated for each iteration, with fewer new judgments required for each iteration
- Metrics: precision, recall, F score at top 5 and top 10 returns

Matching Evaluation Results

	Original Thresholds		Top 10 F	Top 5 F
	Recall	Precision	Score	Score
1. (2) U (3)	.919	.519	n/a	n/a
2. Lx and culture informed	.812	.755	.777	.759
3. Phonetic	.914	.515	.646	.655
4. (2) tuned for precision	.696	.814	.750	.736
5. Arabic-tuned with keys	.866	.631	.739	.747
6. Exclusively for Arabic	.713	.831	.765	.749
7. General fuzzy matcher	.717	.772	.742	.731
8. General record match	.747	.747	.746	.734
9. (8) U (2)	.782	.744	.760	.743
10. Variant Generation	.649	.834	.729	.714
11. Variant Generation	.549	.892	.675	.658
12. Simple Variant Generation	.516	.837	.626	.608

Possibilities for Incorporation of Name Tagging and MT

- <DO_NOT_TRANSLATE> tags
- Named Entity tags
  - <ENAMEX...>
- <TRANSLATE> and <TRANSLITERATE> tags
- Babych & Hartley (2003)

Improving MT with Extraction Technology

- Babych & Hartley (2003) used entity tagger to identify organization names in MUC-6 texts
- Produced "Do Not Translate" lists for English-Russian ProMT 98, English French ProMT, English-French Systran
- Compared translations with and without using the "Do Not Translate" lists
- Created scoring system for paragraphs containing the organization names (+1, +0.5, 0, -.5, -1)
- Scored contexts of 50 names and computed a percent improvement: 29%, 22%, 32%
- Issues with DNT list, e.g. Labour party vs. common noun

Extraction Empowered MT

- Use extraction software to identify names
- Treat names as generalized tokens in second-order phrase translation rules: a special case of hierarchical phrasal translation
- Can write transfer rules for core NP patterns
- Process names with a hybrid model consisting of
  - Standard statistical MT phrase translation
  - Transliteration hypotheses for person-names
  - Type-specific lexicons where available
  - Combine scores with weighted interpolation
- Preliminary tests did not produce large BLEU score increases, but improvements in MT quality were evident
  - Will talk about this more in evaluation section

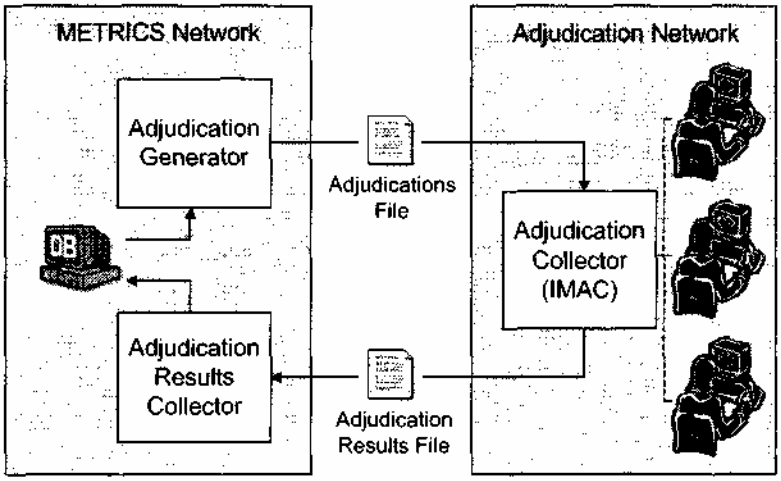
Evaluation: Names and MT

- What constitutes a good translation / transliteration of a name?
  - What constitutes a valid variant of a name?
- What constitutes a good evaluation
  - for names?
  - for MT?
- What constitutes good ground truth data?
- What metrics are sensitive to proper translation of names in MT?

Development of Ground Truth Sets

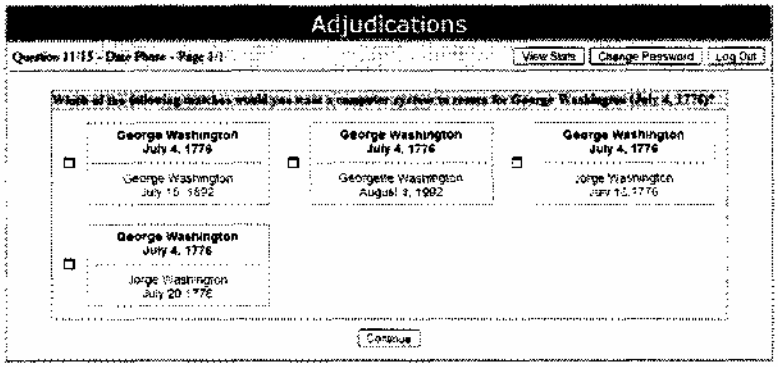
- Usual issues of representation apply
- Human adjudication required, but need not be exhaustive
- Some available resources ("equivalent" surnames in *Family History Knowledge UK* [Park, 1992])
- Generate variants
  - Human generation
  - Automatic generation
- Estimate truth
  - Pooled methods using multiple matching methods
  - Each method returns n match candidates for human adjudication
  - Combine results and adjudicate top 200

Adjudication Collection



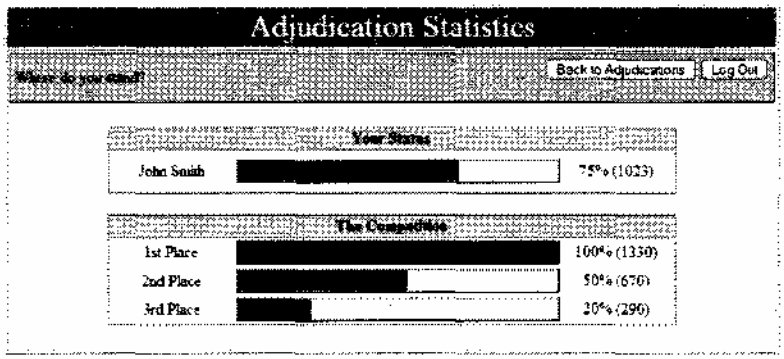
IMAC - User Interface

One screen of the Adjudication Collector continually provides questions to the adjudicator which need to be answered. These screens first ask the question with no dates provided and then again asks the question with dates shown.



IMAC - User Interface (2)

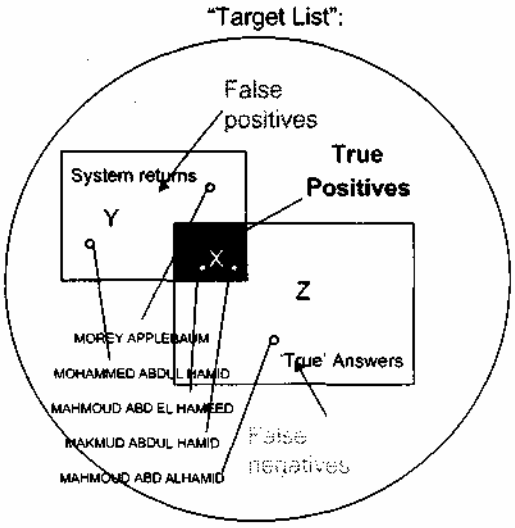
Another screen of the Adjudication Collector shows how the adjudicator is doing compared to the others. A reward of some sort might be provided to whomever completes the most adjudications.



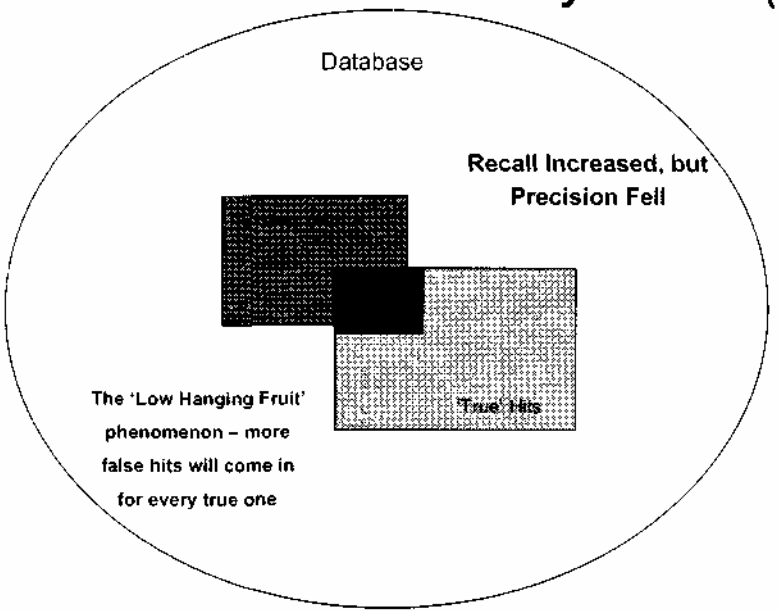
Basic Metrics: Precision and Recall

"Subject":
MAHMOUD ABDUL HAMEED
12/10/1945

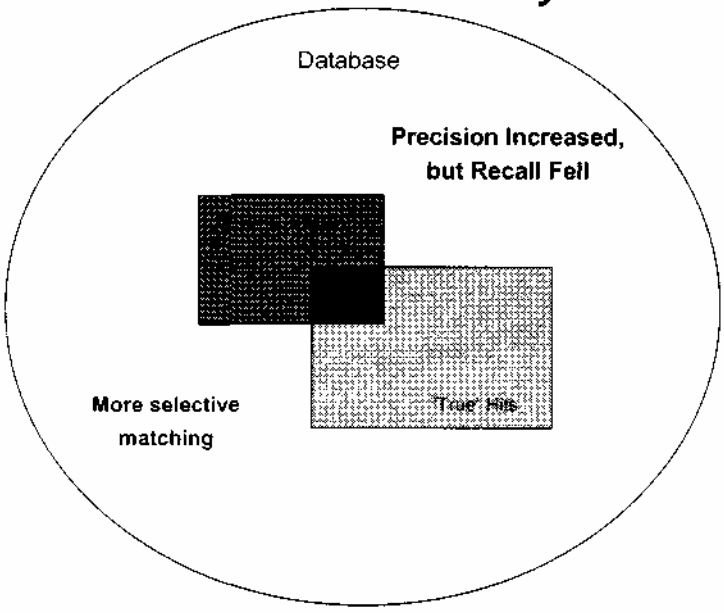
Precision (P) = X/Y (2/4)
Recall (R) = X/Z (2/3)



Precision and Recall Inversely Related (1)



Precision and Recall Inversely Related (2)



Evaluation: Like IR Tasks

- Metrics
  - F-measure - harmonic mean of precision and recall
    - $F = (\beta^2 + 1) P R / ((\beta^2 P) + R)$ where
      - P = precision = correct system responses / all system responses
      - R = recall = correct system responses / all correct reference responses
      - β = beta factor– provides a mean to control the importance of recall over precision
  - Additional Measures
    - False positives – items that are identified as correct responses that are not correct responses
      - (= 1 – Precision)
    - False negatives – correct responses not identified
      - (= 1 – Recall)
    - Fallout = non-relevant responses / all non-relevant reference responses (related to, but not directly calculable from precision / recall)
  - Issue:
    - Annotation Standard for Development of Ground Truth

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
- Well-documented – also contextualizes results in terms of purpose of the evaluation and task

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

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

Originally developed as an evaluation framework for Machine Translation, but authors note that it should be able to be used as a generic evaluation framework.

Issues In Establishing Ground Truth

- Different truth for different applications
  - Credit Check
  - Security Applications
  - Customer Support
  - Deduplication of Mailing Lists
- What is the cost of missing a match?
  - New record entered into database
  - Irritated customer
  - Lives are lost
- Criteria for truth must be carefully established and well-understood by annotators
  - Question posed to annotators must be carefully phrased

Issues In Establishing Ground Truth

- How much time / expertise is available to judge (/discount) false positives?
- Evaluation results are only as good as the truth on which they are based
  - And only as appropriate as the evaluation is to the task that will be performed with the operational system
- Absolute recall impossible to measure without completely known test set
  - Estimate with pooled results

Sidebar: Issues In Establishing Ground Truth for Identity Matching

- For identity matching: what identity elements are available:
  - Name only
  - Date of birth
  - Country of birth, citizenship, passport?
  - Passport, driver's license, social security, other ID number
  - Biometrics
  - Other....

B Smith ⇔ Bill Smythe ⇔ William Smythe ⇔ W Smith ??

DOB: 10/12/1972 ⇔ October 11, 1972 ⇔

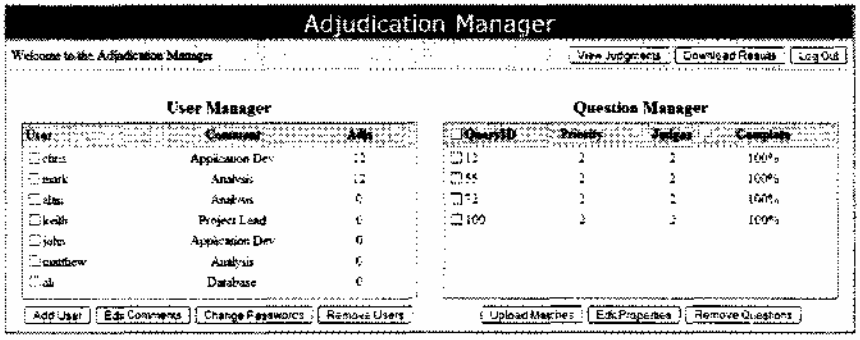
December 10, 1972 ⇔ 12/10/72 ⇔ October 12, 1927

Activity 1: Determining Appropriate Variation in Name Representation

- **See handout**

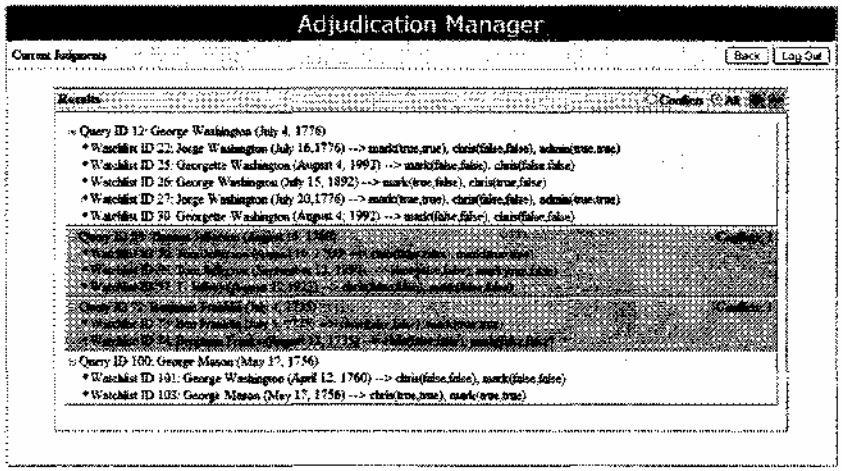
IMAC – Admin Interface

An administrative screen allows the ability to manage IMAC users as well as manage the questions asked of users. This includes the ability to set the priority of questions and the number of judges to be used for each question.



IMAC – Admin Interface (2)

Viewing and resolution of conflicting adjudications can also be performed from the administrative screen.

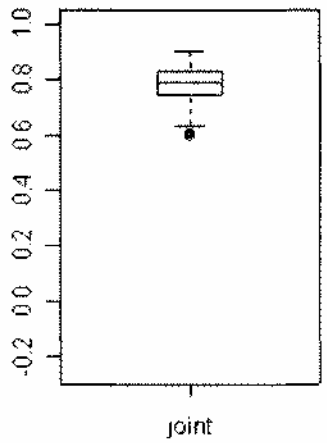


Inter-Assessor Agreement (1)

Joint agreement:

$$\sum_{i=\min}^{\max}\sum_{j=\min}^{\max} p_{i,j}$$

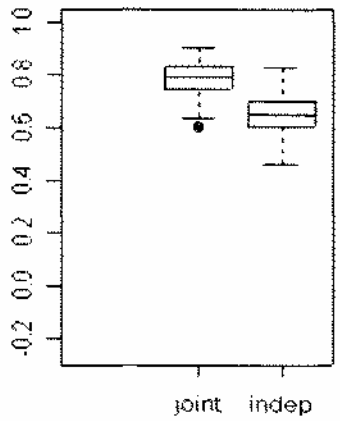
$p_{i,j}$ = proportion of observations in the cell at row i, column j



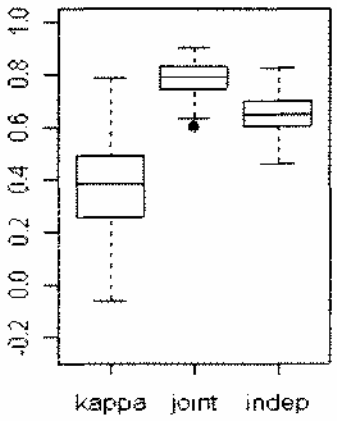
Inter-Assessor Agreement (2)

$$\text{indep\_sum} = \sum_{i=\min}^{\max} \sum_{j=\min}^{\max} p_{i\bullet} p_{\bullet j}$$

$$\text{kappa} = \frac{\text{joint\_sum} - \text{indep\_sum}}{1 - \text{indep\_sum}}$$



Inter-Assessor Agreement (3)



Kappa low due to high independent probability of agreement.

Methods of addressing

- independent
- joint
- statistic

Other issue: determining independent probability

Variant Generation for Ground Truth

- **2 native speakers of Arabic generated 2,800 variant database names (each associated with query name)**
- **Created several variants of query name**

  Muhammad Abd al Amir

  → Mohamed Abdel Amer, Mhammad Abdul Amir
- **Created one "false positive" for each**

  (e.g., Mahmud Abdul Amir)
- **For each query, database contained at least 3 pre-judged matching names and 1 similar but false match for initial ground truth**

  independent of pooled matching results

MT Evaluation Metrics and Proper Names

- **Holistic vs discrete point evaluation**
- **Metrics sensitive to names / metrics not sensitive to names**
- **Possibility of correlation of holistic evaluation with metrics sensitive to proper name handling**
- **Task-based evaluation metrics**

Activity 2: MT Evaluation and Proper Names

- See handout

Performance and Other Considerations

- Speed
- Size of deployment (platform) / memory footprint:
  - room-size
  - mini, PC, handheld
  - server farm....
- Scalability
- Configurability: user dictionaries, domain dictionaries, speed/quality tradeoffs, etc.
- Embedability: APIs (ease of use, granularity)
- Robustness