

Name standardization

(experiments with name transliteration in WIPO)

2 May 2019

Bruno Pouliquen
Global Database Division

Introduction

“Advanced Technology Applications Centre”

Explore machine learning techniques in the IP domain

- Text:
 - Neural Machine Translation
 - Automatic classification
- Speech
- Image
- ...

Is transliteration a simple task?

... does it impact name standardization?

- معمر القذافي(ar,ps)
- Муамар Каддафи (bg,ru)
- Mouammar Kadhafi (Eu,rw)
- Muammar Gaddafi (Eu,sw)
- Moammar Gadhafi (da,sw)
- Muammar al-Gaddafi (Eu,yo)
- Muammar Kadhafi (en,ro)
- معمر قذافي(fa)
- Моамар Кадафи (bg)
- معمر القذافي(fa,ur)
- Муамар Кадафи (bg)
- Muammar Gheddafi (it)
- ...

100 orthographies ;-)

معمر القذافي(ar,ps) Moamer Gaddafi (da,sw)
Муамар Каддафи (bg,ru) Moammar al-Kadhafi (nl)
Mouammar Kadhafi (Eu,rw) Moammar Qaddafi (en)
Muammar Gaddafi (Eu,sw) Muammar al-Qadhafi (de,sv)
Moammar Gadhafi (da,sw) Muammar Ghadhafi (de)
Muammar al-Gaddafi (Eu,yo) Moammar Kadafi (en,pt)
Muammar Kadhafi (en,ro) Muammar Qadhafi (de,sw)
معمر قذافي(fa) Moamar Gadafi (es,sl)
Моамар Кадафи (bg) Muammer Gaddafi (da,sw)
معمر القذافي(fa,ur) Mouammar El Kadhafi (fr)
Муамар Кадафи (bg) Moamar Kadhafi (fr,pt)
Muammar Gheddafi (it) Muammer Gadaffi (Eu,en)
Muammar Kaddafi (da,tr) Muhammar Gaddafi (de,sv)
Muammer Kaddafi (da,tr) Moammar Gheddafi (it)
Muamar Gadafi (es) Muammar al-Qaddafi (Eu,jv)
Moamerja Gadafija (sl) Muammar al-Kadhafi (fr,nl)
Moamar Gaddafi (da,sw) Mummar Gaddafi (en,no)
Muamar el Gadafi (en,pt) Muammar Gadaffi (en,sv)
Muamar Kadafi (eo,pt) Moammar Ghadafi (en)
Muammar Kadafi (en,pt) Moammar al-Qadhafi (de,nl)
Muammar Khadaffi (es,sv) Moammar Khadaffi (nl,sv)
Moammar Gaddafi (da,sw) Mouamar Khadafi (en,pt)
Moamer Kadhafi (en,nl) Moamar Kadhafi (en,pt)
Moammar Kadhafi (fr,sv) Moamer Gadhafi (en)
Muammar Gadafi (en,sw) Muammar Ghadaffi (en,sv)
Muammar el Gaddafi (Eu,es) Moammar Khadafi (fr,pt)
Moamer Gadafi (sh,sl) Muammar Al Gaddafi (de,ms)
Muammar Gadaffi (da,sw) Moeammar Kadhafi (nl)
Mouammar Khadafi (fr,pt) Moamer Kadaffi (nl)
Муаммар Каддафи (os,ru) Moamer Qadhafi (en)
Moammer Kadhafi (en,nl) Muammar el-Gaddafi (de,en)
Muammar Qaddafi (en,sw) Muammar al-Qathafi (en)
Muammar Khadafi (en,ro) Muammar Al-Qadhafi (en,sw)
Muamar Kadhafi (es,pt) Muammar Ghadafi (en,sw)
Muammar al Gaddafi (de,hr) Moammer Gaddafi (en,sw)
Muamar al Gadafi (en,pt) Moammer Gadhafi (en)
Muammar Khadaffi (nl,sv) Muamar El Gadafi (es)
Muhammar Gheddafi (it) Muammar Al-Gaddafi (de,no)
Moamer Gadhafi (en) Muammar el Gadafi (es)
Muammar el-Qaddafi (en,pt) Muamerja Gadafija (sl)
Muammar Gadhafi (de,sw) Moammar Kaddafi (fr,pt)
Moammar Gadaffi (da,sv) Muhammar Kadhafi (pt)
Mouammar Kaddafi (fr,nl) Muammar al-Gadhafi (de,ro)
Muammar Kadaffi (nl,sv) Mouammar Al Kadhafi (fr)
Muamar Khadafi (es,pt) Muammar Khadafy (es,pt)
Muammar al-Ghadhafi (de) Moammar Khadaffi (nl,sv)
Muammar Gadafy (en) Moammar Khaddafy (en)
Muamar Gaddafi (da,sw) Muammar Kadhaffi (pt)
Muammar al-Gadafi (es,no) Muammar Al-Kaddafi (pl)
Moammar Qadhafi (de,en)

Source: <http://emm.newsexplorer.eu/NewsExplorer/entities/fr/262.html>

Transliteration: challenges

- E.g. application SU27926309



Государственный комитет
СССР
по делам изобретений
и открытий

К ПАТЕНТУ

(61) Дополнительный к патенту —	(21) 2302898/18-25	(51) М. Кл. ²
(22) Заявлено 26.12.75	(32) 30.12.74	G 21 C 13/10
(23) Приоритет —	(33) США	
(31) 537517	Опубликовано 25.01.80. Бюллетень № 3	(53) УДК 621.039
	Дата опубликования описания 05.02.80	.5(088.8)

(72) Автор

Иностранец
Джек Е. Джонсон
(США)

(71) Заявитель

Иностранная фирма
«Вестингхауз Электрик Корпорейшн»
(США)

Иностранец
Джек Е. Джонсон
(США)

TS OF EXOTHERMAL ASSI

(54) УСТРОЙСТВО ДЛЯ УЛАВЛИВАНИЯ РАСПЛАВЛЕННОГО
ТОПЛИВА И ОБЛОМКОВ КОНСТРУКЦИИ ТЕЛОВЫДЕЛЯЮЩИХ

PermaLink

Office : Russian Federation(USSR data)
Application Number: 2302898 Application Date: 26.12.1975
Publication Number: 00712050 Publication Date: 25.01.1980
Publication Kind : A3
IPC: 5G 21C ?
Applicants:
Inventors: DZHEK E. DZHONSON
Priority Data: 74 537517 30.12.1974 US
Title: (RU) Устройство для улавливания расплавленного топлива и обломков конструкции тепловыделяющих сборок ядерного реактора
(EN) DEVICE FOR CAPTURING MOLTEN AND STRUCTURAL FRAGMENTS OF EXOTHERMAL ASSEMBLIES OF NUCLEAR REACTOR

ДЖЕК Е. ДЖОНСОН

DZHEK E. DZHONSON

Transliteration, back-transliteration

- Transliteration:

- Jack E. Johnson → ДЖЕК Е. ДЖОНСОН

- Back-transliteration:


- (good) ДЖЕК Е. ДЖОНСОН → Jack E. Johnson
- (bad) ДЖЕК Е. ДЖОНСОН → DZHEK E. DZHONSON

Future/current work on transliteration already on Patentscope (currently only for Japanese)

Application Number: 2006551087 **Application Date:** 20.12.2004

Publication Number: 2007520013 **Publication Date:** 19.07.2007

Publication Kind : A5

IPC: G07F 9/10 

Applicants: ザ コカコーラ カンパニー

Coca Cola Co

Inventors: ラディック, アーサー ジー,
アンタオ, レオナード エフ,

Radic Arthur G
aAtao Leonard F

Agents: 山本 秀策
安村 高明
森下 夏樹

Shusaku Yamamoto
Yasumura Takaaki
Natsuki Morishita

Priority Data: 10/708,005 02.0

Title: (JA) 温冷両用自動販売機

Abstract: (JA)

温冷両用自動販売機。この自動販売機は、製品用区画(141)、冷蔵システム(305)、ならびにこの冷蔵システムおよびこの製品用区画と連通する通風システム(180)を備え得る。通風システムは、この製品用区画と連通するように配置されたバルブ(240)を備え得る。ヒーター(270)が、この製品用区画のまわりに配置され得る。このバルブおよびこのヒーターは、この製品用区画が暖められるかまたは冷却され得るように、選択的に活性化される。

Name transliteration

- Character based sequence-to-sequence monotonic decoding
- NMT “learns” transliteration rules from “noisy” examples

Аанеста́д Лейф Инге	Aanestad Leif I. (Inge)
ААН ХЕНДРИК КЛАЗИНГА	KLASINGA AAN HENDRIK
АГРЕ Дэниел Х.	D. H. Agre
Альбрехт Эденхофер	dr Albrecht Edenhofer

Prototype:

Transliteration

Source text:

Result: morishita natsuki

فيليب ماجنير

Philippe Magner

карлос абад

Carlos Abad

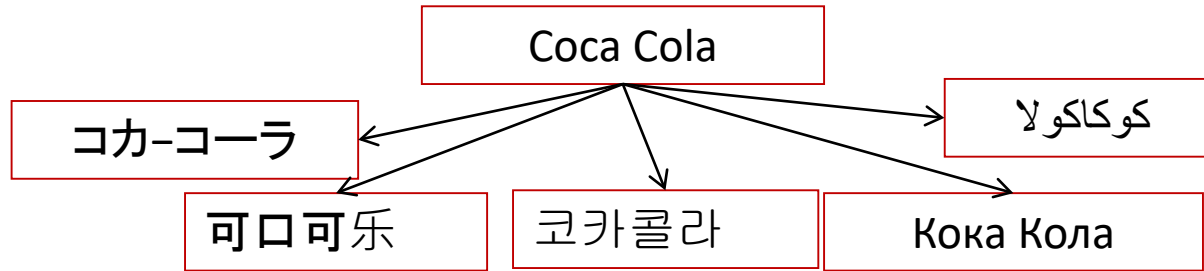
鲍里斯·布里亚肖夫

Boris Blyashov

Challenges:



Proper name transliteration



Demo

- Similar to machine translation
 - “translation” from/to different scripts
- “learn” from existing transliterations
- “guess” possible spelling in different script
 - LI, Lili => 李丽丽
 - Francis Gurry=>弗朗西斯·古里
 - 可口可乐 => Coca-Cola

Applicants:	Boditechmed Inc. 韩国帕克特生物科技有限公司
Inventors:	Choi Euiyeol 崔义烈 Nam Kibong 南基凤 Kim Jaehoon 金载勳 Jung Dongseok 郑东锡 Park Sangyeol 朴相烈 Moon Jungdae 文政大 Jung Jinha 郑晋河 Kim Youngmin 金永敏 Jung Soyoung 郑素映 Park Aekyung 朴爱京 Kim Byoungchul 金丙澈 Kim Sungjoong 金成中

Other examples on using AI on proper names

- “learn” from patent applicant/country
- “guess” country of a person name:
 - HAN, Guydon => KR
 - CHEN, Lili => CN
 - Viswanathan, Anand=>IN
 - Wojtaszek, Radoslaw=>PL
 - ...
- Proper name classification (company or person?):
 - Metal Paris => Company
 - Paris Overton => Person



Demo

Early prototype available (English only)

Applicant name standardization (future work)

- Use machine learning
- “learn” from original-name/standardized-name
- “guess” standardized version of a new name
 - Apple => APPLE, INC.
 - Applle, inc => APPLE, INC.
 - ...
- Get access to training data
- Machine learning alone? Hybrid (combined with rules)? human supervision?

Discussion/future work

- This is exploratory work ;-)
- Future includes:
 - Build complete transliteration models:
 - Latin
 - Korean (Hangul)
 - Chinese
 - Japanese (Katakana+Hiragana+Kanji)
 - Arabic
 - Cyrillic
 - Integrate transliteration on Patentscope
 - Gather more/cleaner training examples
 - Address the Neural Machine Translation specific problems with proper names