Phonetic Search in Foreign Texts

Iveta Mrázová, František Mráz, Martin Petříček, Zuzana Reitermanová

Department of Computer Science
Faculty of Mathematics and Physics

Charles University in Prague, Czech Republic

ANNIE 2008, St. Louis, USA – November 11, 2008

Outline

Introduction

Motivation: A foreigner visiting the U.S.

Introduction

Example (Czech → Arabic)

Heard	Arabic	Pronunciation	Meaning
kalbun ⇨	كلب	[kalbun]	dog
	قلب	[$\underline{\text{k}}$albun]	heart

Speaking the same language: An Iraqi-American translator asks census questions in Khandari, Iraq, just west of Baghdad in 2006. JACOB SILBERBERG/AP/FILE

U.S. ARMY HOPES TO KEEP NATIVE ARABIC SPEAKERS

Incentives likely to include large payments to soldiers now working as translators.

By Gordon Lubold, Staff writer of The Christian Science Monitor

Washington – The army may begin paying a retention bonus of as much as $150,000 to Arabic speaking soldiers in reflection of how critical it has become for the US military to retain native language and cultural know-how in its ranks.

A new problem in information retrieval

Inputs:

- phonetic transcription of a word as heard by a foreigner
- large-scale collection of texts / index

Problem:

1. **Which words have the same or similar pronunciation?**
2. **Search the texts for all such words!**

Alternative means

Non-native automatic speech recognition systems (ARS)

+ sophisticated recognition of phonemes
– lower performance than for native speech
– difficulty in hearing and pronouncing all phonemes

International Phonetic Alphabet (IPA)

+ standardized representation of spoken language
– complicated for standard users (tourists,...)

Phonetic algorithms

- code words by their pronunciation
- assign the same code to all spelling variants of the same name (e.g. *Smith*, *Smithe* and *Smyth*)

Phonetic algorithms

Soundex

- words coded by a letter and three digits, eg. R163 for Robert
- \+ simple algorithm with good results for English names
- – many false-positives and false-negatives
- – good performance only for names

English Soundex table

Code	Letters
1	b, f, p, v
2	c, g, j, k, q, s, x, z
3	d, t
4	l
5	m, n
6	r

Phonetic algorithms

Soundex variants - for English:

- *Phonix, Metaphone, NYSSIS,...*

for German:

- *D-M Soundex, Cologne phonetic, PHONEM,...*

for Arabic:

- *Arabic Soundex, Arabic Phonix*
- target English names in Arabic texts

Arabic Soundex

(a) Arabic Soundex table to code the initial letter

Arabic	ا	ب	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ي
Latin	A	B	T	T	J	H	K	D	Z	R	Z	S	S	S	D	T	Z	A	G	F	Q	K	L	M	N	H	W	Y

(b) Arabic Soundex table to code the rest of the word

Code	Letters
omit	ا, و, ي, ع, ح, ه
1	ف, ب
2	خ, ج, ز, س, ص, ظ, ق, ك, غ, ش
3	ت, ث, د, ذ, ض, ط, ة
4	ل
5	ن, م
6	ر

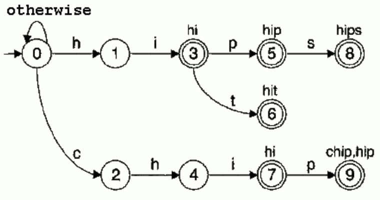
Our approach to phonetic search - CZFind

- the main idea opposite to phonetic algorithms (Which words have the same or similar pronunciation?)
- language-dependent transcription rules and pre-processing
- heard words are searched with purpose-generated Aho-Corasick Automata
- improved speed and precision with dictionaries

Aho-Corasick Automaton

A finite state machine, that searches for all occurrences of a finite set of strings.

- tree-like structure
- linear complexity



state	1	2	3	4	5	6	7	8	9
failure state	0	0	0	1	0	0	3	0	5

Supporting experiments

CZFind implemented for:

1. **Czech → German**
2. **Czech → Arabic**

Data:

- **German**: 2300 articles from the German Wikipedia (23.4 MB), a German dictionary (500000 words), 124 randomly selected target words with a "Czech-like" pronunciation (Kühlschrank → kýlšrank)
- **Arabic**: 950 articles from the Arabic Wikipedia (6.8 MB), an Arabic dictionary (30000 words, Sameer), 36 randomly selected target words with a "Czech-like" pronunciation

Czech-German

Pre-processing

- conversion of letters to lower case
- replace multiple letters by just one occurrence

Czech-German rewriting rules

Czech	German
a, á	a
ä, e, é	ä
b	b
c	c, tz
d, t	d
e, é	e
f	f, pf, ph
g, k	g
h	h
i, í, y, ý	i, ü
j	j, i
k	c, k, ck, ch
l	l, el
m	m
n	n, ng, en
o, ó	o

Czech	German
ö, é, e	ö
p	p
q, kv	q
r	r, er
s, z	s
t, th	t, dt, th
u, ú, ů	u
ü	ü
v, f	v
v, w	w
x, ks	x
i, í, y, ý, j	y
z, c	z
ß, s, ss	ß, ss
oj	eu, äu
ai, aj	ei, ai, ay, ey

Czech	German
š	sch, ch
šp	sp
št	st
č	tsch, tzsch
kv	qu
ich, ik	ig
ks	chs
a, á	aa, ah
é, e	ee, eh, oe
é, e	äh, öh
ä	äh
ö	öh
ü	üh
í, ý	ie, üh
o, ó	oo, oh
u, ú, ů	uh, uu

Czech-Arabic

Pre-processing

- decomposition of ligatures
- conversion of letters to their general form
- removal of some characters (Shadda, Hamza)

Examples of Czech-Arabic rewriting rules

Czech	Arabic
á, a, i, áj	ا
aj, ajá, íjá, íja, íjá	يا
b, p	ب
t	ت
th	ث
j, g, ž, č, dž	ج
h, ch	ح
k, kh, x, ch	خ
...	...

Czech	Arabic
s	س
š, sh	ش
s	ص
d	ض
t	ط
z	ظ
r,ch	ع
gh, g, h, r, ch, chr	غ
f	ف
q, k	ق
...	...

Czech	Arabic
n, m	ن
h	ه
v, w, ů, ú, u	و
y, i, j, í, ý, íj, á	ي
x	كس
a, e, i, o, u, y, ý, í	*(empty)*
č	تش
un	ٌ
a	َ
...	...

Comparison of CZFind and phonetic algorithms

1. How many codes cover all words accepted by the ACA?
2. How many words from the dictionary get the same code?

Algorithm	**Number of distinct codes**			**Number of dictionary words with the same code**		
	average	min	max	average	min	max
German						
Cologne Phonetic	1.10	1	2	85.4	1	296
PHONEM	1.17	1	3	10.3	1	77
Soundex	1.26	1	3	200.4	7	1037
Daitch Mokotoff	1.38	1	3	17.0	1	85
Arabic						
Arabic Soundex	1.81	1	5	733.6	6	2316
Arabic Phonix	2.33	1	10	512.4	1	1955

Precision of retrieval - German

1. How many words will be retrieved?
2. How many of the retrieved words will be correct?

Algorithm	CZFind	PHONEM	DM Soundex	Soundex	Cologne
Average number of distinct words retrieved from the text (over 124 words)					
correct	1.8	2.2	1.2	8.6	2.9
all	5.6	14.1	16.8	121.4	112.8
correct ratio	0.53	0.44	0.30	0.08	0.11
Average number of all words retrieved from the text (over 124 words)					
correct	1695.8	798.9	745.8	600.4	570.6
all	2521.0	1381.5	2218.8	2698.6	6171.1
correct ratio	0.79	0.74	0.62	0.28	0.33

Precision of retrieval - Arabic

1. How many words will be retrieved?
2. How many of the retrieved words will be correct?

Algorithm	CZFind	Arabic Soundex	Arabic Phonix
Average number of distinct words retrieved from the text (over 36 words)			
correct	1.8	6.3	4.7
all	6.9	248.4	302.5
correct ratio	0.53	0.09	0.05
Average number of all words retrieved from the text (over 36 searched words)			
correct	205.0	93.1	67.1
all	369.6	1410.4	1656.8
correct ratio	0.70	0.17	0.09

How fast are the algorithms?

	Initialization time			Search time		
	Average	**Min**	**Max**	**Average**	**Min**	**Max**
Cologne Phon.	0 s	0 s	0 s	0.36 s	0.33 s	0.42 s
Regular expr.	0.019 s	0.017 s	0.05 s	9.11 s	4.05 s	127.02 s
Aho-Corasick	0.022 s	0.006 s	1.37 s	2.19 s	1.83 s	2.60 s

Conclusions

CZFind $\sim$ a quick, precise and user-friendly approach to phonetic search

- A viable solution to a new problem in information retrieval
- Retrieval precision comparable with the best German algorithms and $4\times$ better than Arabic algorithms
- Significantly faster than regular expressions for large text collections or indexes
- **Adds semantics to retrieved documents**

Further research

- Automatic learning of rewriting rules