شماره مستند: ۱۹۰/۲۵۳۷/۱/۸

جمهوری اسلامی ایران

دبیرخانه شورای عالی اطلاع رسانی

تحلیل نیازمندی‌های سیستم تشخیص کلمات فینگلیش در لابه‌لای متون زبان فارسی

نسخه ۱.۰

دانشگاه علم و صنعت ایران

فروردین ۸۸

فهرست

۱ مقدمه

پیش از همه‌گیر شدن استفاده از چیدمان فارسیِ صفحه‌کلید توسط فارسی زبانان، رسم‌الخطی به نام پینگلیش بوجود آمده بود که در آن از کاراکترهای انگلیسی برای نوشتن کلمات فارسی استفاده می‌شد. استفاده از این رسم‌الخط در سال‌های اولیه‌ی ورود اینترنت به ایران رواج زیادی داشت، چرا که هنوز استاندارد مشخصی برای تبادل کاراکترها و متون فارسی که اولاً موردقبول اکثریت بوده و ثانیاً در همه‌ی سیستم‌های کامپیوتری قابل دسترس و استفاده باشد، وجود نداشت. امروزه استاندارد یونیکد این مشکل را تاحدود بسیاری مرتفع کرده است اما با این وجود هنوز هم استفاده از رسم‌الخط پینگلیش میان کاربران رایج است. مهم‌ترین حوزه‌ای که امروزه استفاده از پینگلیش در آن رواج دارد، دستگاه‌های قابل حمل ارتباطی مانند PDA و موبایل است. در اینترنت نیز هنوز برخی از کاربران بدلیل عدم آشنایی کامل با چیدمان فارسی صفحه کلید، متون خود را پینگلیش تایپ می‌کنند.

حجم قابل توجهی از متون برخی سایت‌ها نیز (مانند فروم‌ها) پینگلیش می‌باشد. برخلاف نوشتنِ پینگلیش که بنظر ساده می‌آید، خواندن متون پینگلیش (بویژه متونی که بیش از چند پاراگراف باشند) امری خسته‌کننده و وقت‌گیر است. با در دست داشتن ابزاری برای تبدیل متون پینگلیش به فارسی، می‌توان این نقیصه را برطرف نمود.

۲ ساختار متون پینگلیش

بررسی متون پینگلیش نشان می‌دهد که قواعد و قوانین مشخصی برآن حاکم نمی‌باشد. دلیل این امر نیز تاحدودی واضح است: پینگلیش توسط عامه‌ی کاربران فارسی‌زبان ایجاد شده و مرکز یا مبدا مشخصی برای تدوین قواعدِ پینگلیش‌نویسی وجود نداشته است، لذا تنوع زیادی در قواعد و الگوهای پینگلیش‌نویسی وجود دارد.

بعنوان مثال، کلمه‌ی «اعتماد» در پینگلیش ممکن است به شکل‌های زیر نوشته شود:

etemad, etemaad , e’temad, e’temaad, eatemad, …

همانگونه که مشاهده می‌شود، گاهی برای ۱ کلمه‌ی فارسی ممکن است بیش از چندین نوع نحوه‌ی نگارش پینگلیش وجود داشته باشد. یکی از دلایل این چندگانگی وجود داشتن چندین کاراکتر انگلیسی برای برخی از حروف فارسی است.

یکی دیگر از ویژگی‌های متون پینگلیش این است که شامل کلمات و اصطلاحات غیررسمی بوده و کلمات در این متون عموماً بصورت شکسته بیان می‌شوند. دخیل کردنِ حالات و احساسات در بیان کلمات نیز امری متداول در پینگلیش بشمار می‌رود.

۲.۱ *نگاشت حروف فارسی و انگلیسی*

در جدول ۱، حروف فارسی و معادل‌های رایج آن‌ها در پینگلیش به تفکیک آمده‌اند. در برخی خانه‌های ستون دوم جدول (معادل‌های رایج فینگلیش) کاراکتر $ مشاهده می‌شود؛ این کاراکتر برای نمایش کاراکترِ «خالی» مورد استفاده قرار گرفته است. بعنوان مثال برای حرف ع، به نگاشت زیر در کلمه‌ی «اعتماد» دقت کنید:

اعتماد	etemad
ا	e
ع	$
ت	t
ـِ	e
م	m
ا	a
د	d

جدول ۱ - حروف فارسی و معادل‌های آن در پینگلیش

حرف فارسی	معادلهای رایج فینگلیش	متداولترین معادل	مثال
حروف بیصدا			
ب	b	-	bache
پ	p	-	pedar
ت	t	-	tatilat
ث	s, c	-	mosbat
ج	j, g	j	jahat
چ	ch	-	cheghadr
ح	h	-	mohit
خ	x, kh	kh	khanevadeh
د	d	-	dar
ذ	z	-	begzarim
ر	r	-	rah
ز	z	-	zir
ژ	zh, j	j	mojdeh
س	s,c	s	saket
ش	sh	-	shabih
ص	s,c	s	saboor
ض	z	-	zaroori
ط	t	-	tathir
ظ	z	-	zaher
ع	a, a’, ee, ‘, $		moeen, mo’een
غ	gh, q	-	ghalat, qalat
ف	f,ph	f	farda
ق	gh, q	-	ghader
ک	k,c	k	kah
گ	g	-	gom
ل	l	-	lazem
م	m	-	mamnoon
ن	n	-	naan
ه	h, $	h	hadyeh, had
حروف باصدا			
آ	a, aa, $	a	aab
ا	a, aa, e, $	a	
ی	i, y, ei, ie, ee, e, iy, ey,	y,i	yeki, ieki, yeky

	ye, yi		
ُ(ضمه)	o, $	o	mohit
و	o, oo, ou, uo, v, u, w	o,oo,v	zood, va'de, wared
ِ(کسره)	e, $	e	mehman
َ(فتحه)	a, $	a	mashroot
حروف و کاراکترهای غیررسمی			
سی	C, 3o	-	merC, mer3o
سه	۳	-	madre3
عت	@	-	sa@
َت	@	-	
تی	T	-	rafT
بی	B	-	charB
دی	D	-	boD hala
زی	Z	-	Zafati (ضیافتی)
شما	U (در حالتیکه بعنوان یک کلمه بکار رود)	-	
من	I (در حالتیکه بعنوان یک کلمه بکار رود)	-	
اس	S	-	Smaeel
جی	G	-	Geegar
	X	-	
ان	N	-	Nhedam (انهدام)
ام	M	-	Mkanat (امکانات)

همانگونه که در جدول ۱ نیز مشاهده می‌شود، گوناگونی و تنوع بسیاری در نحوه‌ی تبدیل کلماتِ فارسی به معادل پینگلیش آن‌ها وجود دارد. در جدول فوق تلاش شده است تا پرکاربردترین این قواعد آورده شوند. با استفاده از جدول ۱، می‌توان جدول دیگری برای نگاشت از حروف انگلیسی به حروفِ فارسی ایجاد کرد:

جدول ۲- نگاشت حروف انگلیسی به حروف فارسی

حرف انگلیسی	معادل محتمل	متداولترین معادل
حروف با یک معادل		
b	ب	-
d	د	-
r	ر	-
f	ف	-

l	ل	-
m	م	-
n	ن	-
v	و	-
w	و	-
y	ی	
‘	ع	
حروف با چندین معادل		
h	ه ح	ه
x	خ، کس،	خ
t	ت ط	ت
T	تمامی حالات t + تی	ت
s	ث س ص ش = sh	س
S	تمامی حالات s + اس	س
c	ث س ص ک چ = ch	س
C	تمامی حالات c + سی	س
p	ف= ph پ = p	پ
P	تمامی حالات p + پی	پ
j	ژ ج	ژ
g	ج	ج

	ق = gh غ = gh گ	
z	ز ذ ض ظ ژ = zh	ز
a	ـَ(فتحه) ا آ أ عا أ – an ع ی ه وا آ = aa	
e	ـِ(کسره) ی ا ه = eh ع = ee ی = ei ی = ee ی = ey اع	ـِ(کسره) ی
i	ی ای ی = ie ی = iy	ی
k	ک خ = kh	ک
o	ـُ(ضمه) و او و = oo و = ou	و

u	و او و = uo شما (در حالتیکه یک کلمه‌ی یک حرفی مشاهده کنیم)	و
q	ق غ	ی

همانگونه که در جدول ۲ براحتی قابل مشاهده است، فقط تعداد بسیار کمی از حروف انگلیسی، دارای یک معادل در فارسی هستند و بسیاری از حروف انگلیسی، دارای ۲ یا چند معادل در فارسی هستند. همین امر مشکلات بسیاری را در توسعه‌ی یک مبدل پینگلیش ایجاد می‌کند.

۲.۲ تکرار حروف

در متون پینگلیش، تکرار حروف معمولاً به ۲ دلیل رخ می‌دهد:

۱. برای بیان حروفی مانند «و»، بعنوان مثال: mash**oo**t

۲. برای بیان احساسات و ابراز هیجان: merccccccc

علاوه بر ۲ علت فوق؛ ممکن است برای بیان «تشدید» نیز از تکرار حروف استفاده شود، اما بررسی متون پینگلیش نشان می‌دهد که کاربران تقریباً در اکثر موارد، تشدید را در نحوه‌ی نگارشِ کلمات دخالت نداده‌اند. بعنوان مثال املای کلمه‌ی مفرح بصورت «mofarah» رایج‌تر از «mofarrah» می‌باشد.

۲.۳ استفاده از کلمات شکسته

متون پینگلیش عمدتاً بصورت محاوره‌ای نوشته شده‌اند و کلمات در آن‌ها بصورت شکسته بکار رفته‌اند. بعنوان مثال، کلمه‌ی «خانه» معمولاً بصورت khoone نوشته می‌شود و نه khane.

۲.۴ استفاده از کلمات انگلیسی

الگوی دیگری که در متون پینگلیش مشاهده می‌شود، استفاده از برخی کلمات انگلیسی است. این کلمات عمدتاً در دسته‌ی کلمات فنی قرار می‌گیرند و رایج‌ترین حوزه‌ی آن‌ها، حوزه‌ی کلماتِ تخصصیِ کامپیوتر است. برخی از این کلمات در جدول ۳ گردآوری شده‌اند. (لازم بذکر است که این جدول شامل تمامی این کلمات نمی‌باشد.)

جدول ۳ - برخی از واژه‌های انگلیسی رایج در متون پینگلیش

واژه‌ی انگلیسی	معادل
copy	کپی
cd	سی‌دی
site	سایت
page	صفحه
IP	آی‌پی
type	تایپ
check	چک
free	آزاد، خالی
dialup	دایال‌آپ،
comp, computer	کامپیوتر
title	عنوان
CPU	سی‌پی‌یو، پردازنده
memory	حافظه
project	پروژه
mail, email	ایمیل
paste	الصاق کردن
tool	ابزار
hi	سلام
bye	خداحافظ
good	خوب
ok	باشه، بسیارخب،
address	آدرس
tv	تلویزیون

۲.۵ استفاده از حروف و کاراکترهای ویژه در کلمات

از این الگو در برخی موارد برای کوچکتر کردن طول کلمات و خلاصه‌سازی استفاده می‌شود:

جدول ۴

حرف – کاراکتر انگلیسی	معادل	مثال
o	o	zo0d -> zood
1	ye, yek	1doone -> yedoone
2	too, doo, do	be2nin -> betooni, bedoonin 2a- > doa
3	se	3tar -> setar
4	chahar, char, cha, for	4kerim -> chakerim
8	eit	w8-> weit (صبرکن)
@	at	sa@ -> saat
3o	si, ci	mer3o -> merci
T	ti	raft -> rafti
C	si, ci	merC -> merci

۳ مبدل پینگلیش

در بررسی روش‌های متداول تبدیل خودکار متون پینگلیش به فارسی نکته‌ای که جلب توجه می‌کند، عدم توانایی این روش‌ها در پوشش دادن همه‌ی الگوهای ذکر شده در بخش قبل است. برخی از روش‌ها نیز کلمات پینگلیش را درصورتی بدرستی تبدیل می‌کنند که در نوشتن آن کلمه، قوانین خاصی رعایت شده باشد (بطور مثال حرف «ع» بصورت «’» نوشته شده باشد، یا حرف «ا» بصورت «aa» در کلمه‌ی e’temaad).

بنابراین در این پروژه روش جدیدی برای تبدیل متون پینگلیش ایجاد و پیاده‌سازی شد. ایده‌ی اصلی این روش، یادگیری و کشف الگوهای تبدیل توسطِ خود نرم‌افزار است. این روش ۲ مرحله دارد:

۱. **کشف و یادگیریِ الگوهای تبدیل، از طریق تعدادی کلماتِ نمونه**: در این مرحله، تعدادی کلمه‌ی پینگلیش و معادلِ آنها بعنوان ورودی به نرم‌افزار داده می‌شود. نرم‌افزار طبق مراحلی، معادل‌های حروف انگلیسی را یادگرفته و آن‌ها را ذخیره می‌کند.

۲. **استفاده از الگوهای یادگرفته شده در مرحله‌ی قبل**: در این مرحله، کلمات پینگلیش بعنوان ورودی به نرم‌افزار داده می‌شود و نرم‌افزار با استفاده از الگوهای کشف‌شده در مرحله‌ی قبل، معادل‌های محتمل برای آن کلمات را ارائه می‌کند.

یکی از مزیت‌های این روش در این است که می‌توان الگوهای جدید را نیز در هر لحظه به نرم‌افزار آموخت و نیازی به بازنویسی نرم‌افزار نیست.

در ادامه، توضیح هر یک از مراحل فوق می‌آید.

۳.۱ کشف و یادگیریِ الگوهای تبدیل

در این مرحله، تعدادی کلمه‌ی پینگلیش و نحوه‌ی نگاشت آن‌ها به فارسی به نرم‌افزار داده می‌شود. بعنوان مثال ورودی‌های زیر را درنظر بگیرید:

n --> ن
o --> و

a --> ا
m --> م
a --> ا
d --> د
e --> ِ

g --> گ
i --> ی

c --> چ
h -->
e --> ِ
s --> ش
h -->
m --> م
e --> ه

با دردست داشتن نگاشت‌هایی مانند مثال‌های فوق، نرم افزار سعی می‌کند الگوهای نگاشت را کشف کند. بدین‌ترتیب که برای تک‌تک حرف در کلمه‌ی پینگلیش، نگاشت‌های زیر را یافته و ذخیره می‌کند:

- **نگاشت ۲-۲** (۲ حرف قبلی + حرف موردنظر + ۲ حرف بعدی) -» حرف فارسیِ معادل

مثال: حرف e در ch**e**shme

c h **e** s h -> کسره

- **نگاشت ۲-۱** (۱ حرف قبلی + حرف موردنظر + ۲ حرف بعدی -» حرف فارسیِ معادل)
- **نگاشت ۱-۱** (۱ حرف قبلی + حرف موردنظر + ۱ حرف بعدی -» حرف فارسیِ معادل)
- **نگاشت ۱-۰** (حرف موردنظر + ۱ حرف بعدی -» حرف فارسیِ معادل)

مثال: حرف c در **c**heshme

c h -> چ

- **نگاشت ۰-۰** (حرف مورد نظر -» حرف فارسیِ معادل)

۳.۲ استفاده از الگوهای یادگرفته شده

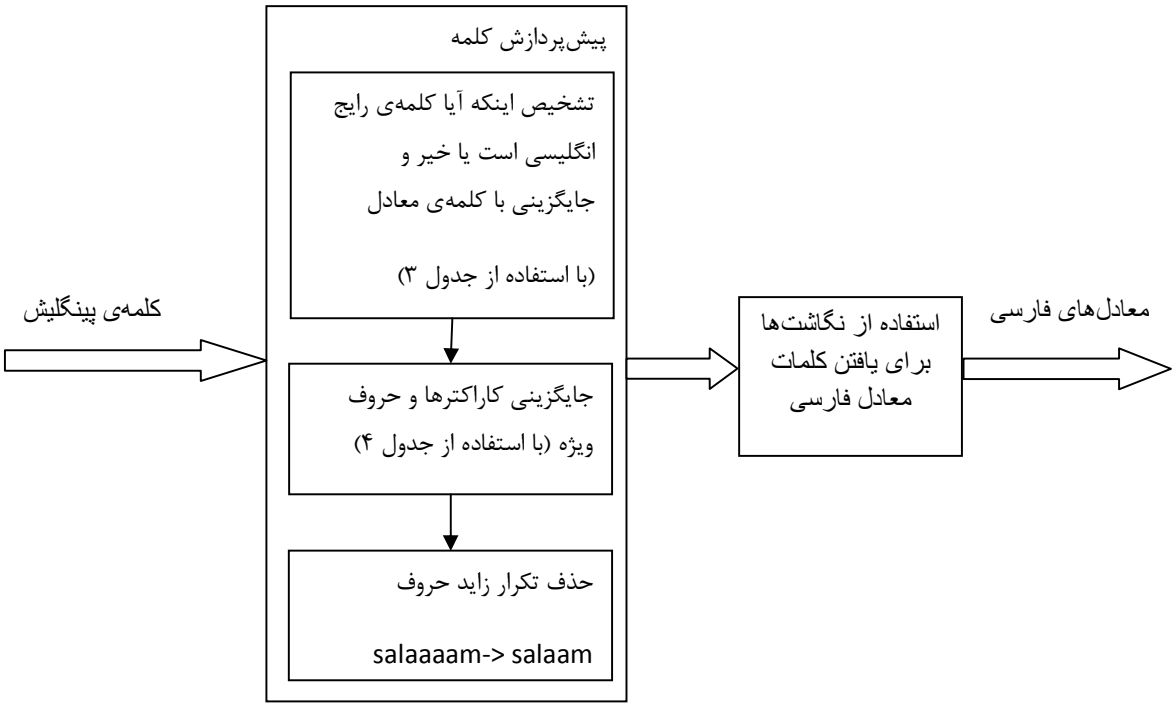
در این مرحله، نرم‌افزار از نگاشت‌های ذخیره شده در مرحله‌ی قبل استفاده می‌کند تا معادل فارسیِ کلماتِ جدید را تولید کند. برای این منظور، الگوریتم زیر را اجرا می‌کند:

۱. برای تمامیِ حروفِ کلمه‌ی پینگلیش، نگاشت‌های ۲-۲، ۲-۱، ۱-۱، ۱-۰ و ۰-۰ را بترتیب جستجو کن. اگر نگاشت مشابهی یافت شد، حرفِ فارسیِ معادل آن را استخراج کن.

اگر نگاشت مشابهی یافت شد، جست‌جو را خاتمه بده و به مرحله‌ی بعد برو.

۲. با استفاده از حروف یافت شده در مرحله‌ی قبل، کلمه (یا کلمات) فارسیِ معادل را تشکیل بده

۳.۳ الگوریتم کلیِ تبدیل کلماتِ پینگلیش



۳.۴ دقت الگوریتم

دقت الگوریتم ارائه شده در این مستند، ارتباط مستقیمی با مجموعه داده‌های اولیه‌ی آن دارد. هرچه تعداد کلماتِ پینگلیش و تنوعِ نگارشیِ آن‌ها در مجموعه داده‌های اولیه بالاتر باشد، دقت الگوریتم در تبدیل کلماتِ جدید بیشتر خواهد بود.

بهمین علت، ابزار جداگانه‌ای برای تولید مجموعه داده‌های اولیه توسعه یافته‌است. این ابزار یک متن پینگلیش را بعنوان ورودی دریافت کرده و به‌ازای تک‌تک کلماتِ موجود در آن، همه‌ی معادل‌های فارسی آن کلمه را تولید می‌کند و به کاربر نمایش می‌دهد. (ایجاد تمامیِ معادل‌های فارسیِ یک کلمه‌ی پینگلیش، با استفاده از جدول ۲ صورت می‌گیرد. بدیهی است که تعداد این معادل‌ها در برخی حالات به چندصد عدد می‌رسد). کاربر از میان کلماتِ نمایش داده‌شده، کلماتِ صحیح را تایید می‌کند و نرم‌افزار بصورت خودکار نگاشت‌های آن کلمه را یافته و ذخیره می‌کند. توسط این ابزار براحتی می‌توان یک مجموعه داده‌ی غنی از کلمات و نگاشت‌های آن‌ها ایجاد کرده و دقت نرم‌افزار را افزایش داد.

البته رسیدن به دقت کامل نیز ممکن نیست، بعنوان مثال می‌توان از کلماتی نام برد که بیش از ۲ معادلِ درست فارسی دارند و تعیین معادل صحیحِ آن‌ها، جز با پردازش معناییِ متن امکان پذیر نیست. madar و dar از جمله‌ی این کلمات هستند که اولی ممکن است «مادر» یا «مدار» بوده و دومی نیز ممکن است «دار» یا «در» باشد.