Pattern Matching dalam Penerjemahan Kata yang Ditulis dengan Huruf Katakana ke dalam Bahasa Inggris

Nabilah Shabrina (13508087)
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
if18087@if.itb.ac.id

Pattern Matching merupakan algoritma yang digunakan untuk mencari suatu teks maupun memvalidasi suatu teks. Pada zaman ini telah banyak berkembang berbagai jenis algoritma patten matching.

Bahasa Jepang menggunakan tiga jenis huruf dalam penulisan bahasanya, yaitu huruf hiragana, katakana, dan kanji. Berbeda dari kedua huruf yang lain, huruf katakana dalam bahasa Jepang digunakan untuk menulis serapan bahasa asing maupun nama orang asing. Hampir semua kata asing dalam bahasa Jepang disadur dari bahasa Inggris, sehingga sebenarnya kita dapat menerka apa arti dari maksud kata tersebut. Dengan menggunakan *Hepburn romanization system* atau pentranslasian huruf katakana menjadi huruf romaji/alfabet, kita bisa mendapatkan arti dari kata tersebut setelah dilakukan proses pattern matching terhadap kata tersebut dengan kata dalam bahasa Inggris. Namun sayangnya huruf dalam bahasa Jepang memiliki keterbatasan yaitu tidak ada konsonan akhir di kata selain huruf 'n'. Oleh karena itu terjadi ketidakmiripan antara pengucapan dalam bahasa Inggris dengan kata yang ditulis dengan huruf katakana. Pada makalah ini akan dibahas bagaimana menyelesaikan permasalahan tersebut sehingga kita bisa mendapatkan arti dari kata dalam bahasa Jepang menjadi bahasa Inggris.

***Index Terms— Hepburn romanization system*, *IPA phoneme*, Katakana, pattern matching,**

I. PENDAHULUAN

Pada zaman ini, pattern matching digunakan secara luas baik untuk mencari suatu teks maupun validasi suatu teks. Ada banyak algoritma yang digunakan dalam pencarian teks. Situs-situs pencarian semacam google pun berkembang pesat.

Di antara bahasa-bahasa di dunia, bahasa Jepang memiliki keunikan sendiri dalam penulisan bahasanya.

1.1 Huruf Hiragana, Katakana, dan Kanji

Pada dasarnya, huruf jepang dibagi menjadi tiga macam, yaitu hiragana, katakana, dan kanji. Huruf-huruf tersebut di pakai oleh masyarakat Jepang dalam sehari-harinya.

Perbedaan dari ketiga jenis huruf ini adalah sebagai berikut.

Pada dasarnya, huruf hiragana dan katakana memiliki bunyi yang sama namun berbeda tulisan. Sementara itu huruf kanji merupakan tulisan yang sangat kompleks dan bunyinya pun cukup banyak mengandung huruf dari rangkaian hiragana.

Huruf hiragana adalah huruf yang memiliki satu kata atau dua kata dalam huruf romawi seperti :

a,i,u,e,o
ka,ki,ku,ke,ko
sa,shi,su,se,so,
ta, chi, tsu, te, to
na, ni, nu, ne, no
ha, hi, hu, he, ho
ma,mi, mu, me, mo
ya, yu, yo
ra, ri, ru, re, ro
wa, wo, n

Huruf hiragana di pakai bila huruf kanji tidak bisa di baca, maka akan di tuliskan huruf hiragananya. Penggabungan huruf tersebut menjadi sebuah kata dan bisa di tulis lewat huruf kanji.

Contoh dari penulisan hiragana adalah sebagai berikut:
おかあさん(okaasan) = ibu
がっこう(gakkou) = sekolah

Huruf katakana memiliki bunyi yang sama dengan hiragana tapi berbeda penulisan. Huruf ini dipakai pada kata penyerapan bahasa asing, atau untuk penulisan nama orang asing.

Contoh penulisan huruf katakana:
ラジオ(rajio) = radio
マガジン(magajin) = *magazine*, majalah

Huruf kanji, merupakan huruf yang kompleks yang bearasal dari Cina. Berbeda dari kedua jenis huruf sebelumnya, satu huruf kanji biasanya melambangkan suatu arti.

Contoh penulisan huruf kanji:
漢字(kanji) = huruf kanji
大学(daigaku) = universitas

1.2 Huruf Katakana untuk Serapan Kata Asing

Huruf katakana berfungsi untuk menuliskan kata-kata serapan dari bahasa asing serta nama-nama orang asing. Di dalam bahasa Jepang, 80% kata serapan asing berasal dari bahasa Inggris. Namun bahasa Jepang memiliki kelemahan dalam menyediakan sumber daya huruf. Huruf katakana (dan juga hiragana) terdiri dari gabungan huruf konsonan dan huruf vokal, sehingga tidak ada huruf mati di akhir kata. Hanya huruf 'n' yang merupakan huruf konsonan tunggal. Hal tersebut menyebabkan sering kali kita sulit menangkap maksud orang Jepang dalam pengucapan bahasa Inggris. Sebagai contoh, bila dalam bahasa Inggris disebut 'ketchup', maka dalam bahasa Jepang disebut 'kechappu'. Bahasa Jepang juga melakukan penyederhanaan dalam ucapan kalimat bahasa Inggris. Misalkan kata 'television' dalam bahasa Inggris, dalam bahasa Jepang disebut 'terebi'. Di bawah ini ditampilkan daftar huruf katakana beserta cara bacanya dengan menggunakan huruf alfabet:

a ア	i イ	u ウ	e エ	o オ										
ka カ	ki キ	ku ク	ke ケ	ko コ		kya キャ			kyu キュ			kyo キョ		
sa サ	si shi シ	su ス	se セ	so ソ		sya sha シャ			syu shu シュ			syo sho ショ		
ta タ	ti chi チ	tu tsu ツ	te テ	to ト		tya cha チャ			tyu chu チュ			tyo cho チョ		
na ナ	ni ニ	nu ヌ	ne ネ	no ノ		nya ニャ			nyu ニュ			nyo ニョ		
ha ハ	hi ヒ	hu fu フ	he ヘ	ho ホ		hya ヒャ			hyu ヒュ			hyo ヒョ		
ma マ	mi ミ	mu ム	me メ	mo モ		mya ミャ			myu ミュ			myo ミョ		
ya ヤ		yu ユ		yo ヨ										
ra ラ	ri リ	ru ル	re レ	ro ロ		rya リャ			ryu リュ			ryo リョ		
wa ワ	wo ヲ	n ン												
ga ガ	gi ギ	gu グ	ge ゲ	go ゴ		gya ギャ			gyu ギュ			gyo ギョ		
za ザ	zi ji ジ	zu ズ	ze ゼ	zo ゾ		zya ja ジャ			zyu ju ジュ			zyo jo ジョ		
da ダ	di zi ヂ	du zu ヅ	de デ	do ド		dya ジャ			dyu ジュ			dyo ジョ		
ba バ	bi ビ	bu ブ	be ベ	bo ボ		bya ビャ			dyu ビュ			dyo ビョ		
pa パ	pi ピ	pu プ	pe ペ	po ポ		pya ピャ			pyu ピュ			pyo ピョ		

Gambar 1. Tabel huruf katakana

Berikut ini merupakan contoh penggunaan huruf katakana dalam penggunaan suatu kalimat dalam bahasa Jepang.

明日テレビでサッカーの試合を見ます。 (1)
Ashita terebi de sakkaa no shiai wo mimasu. (2)
I will watch soccer game on television tomorrow. (3)

Pada kalimat pertama menggunakan huruf hiragana, katakana, dan kanji sekaligus. Penulisan kalimat seperti ini yang umum terdapat dalam bahasa Jepang. Kalimat kedua merupakan cara membaca huruf tersebut yang ditulis dengan menggunakan huruf romaji atau alfabet. Dari kalimat yang digarisbawahi dapat dilihat bahwa kata テレビdibaca *terebi* dan サッカーdibaca *sakaa*. Kalimat ketiga merupakan kalimat yang sudah ditranslasikan ke dalam bahasa Inggris. Kata *terebi* merupukan saduran dari kata *television* dalam bahasa Inggris, dan *sakkaa* merupakan saduran dari kata *soccer* dari bahasa Inggris pula.

Dalam penerjemahan dari huruf katakana menjadi bahasa Inggris, maka dibutuhkan dua langkah. Pertama, mentranslasikan dari huruf katakana menjadi huruf alfabet, kemudian menerjemahkan huruf alfabet tersebut menjadi bahasa Inggris dengan algoritma string matching approximasi. Di sini tidak dibahas mengenai pertranslasian dari katakana ke dalam huruf romaji, namun akan dibahas mengenai penerjemahan katadalam bahasa Jepang yang ditulis dengan huruf romaji ke dalam bahasa Ingris.

II. ALGORITMA PENCOCOKAN STRING

Proses utama dalam penerjemahan kata yang berasal dari huruf katakana ke dalam bahasa Inggris yaitu dengan cara mencari kata asal bahasa Inggris yang mirip dengan kata dalam bahasa Jepang yang ditulis dengan katakana.

Hingga saat ini, telah dikenal berbagai jenis algoritma pencocokan string, seperti algoritma Knuth Morris Path, Boyer Moore, brute force, dan lainnya. Secara umum, algoritma pencocokan string memiliki dua jenis komponen, yaitu

1. teks (text), yaitu (long) string yang panjangnya n karakter
2. pattern, yaitu string dengan panjang m karakter (m < n) yang akan dicari di dalam teks.

Di bawah ini merupakan contoh dari algoritma brute force dalam pattern matching.

```
Algorithm BruteForceMatch(T, P)
    Input text T of size n and pattern
        P of size m
    Output starting index of a
        substring of T equal to P or −1
        if no such substring exists
    for i ← 0 to n − m
        { test shift i of the pattern }
        j ← 0
        while j < m ∧ T[i + j] = P[j]
            j ← j + 1
        if j = m
            return i {match at i}
        else
            break while loop {mismatch}
    return -1 {no match anywhere}
```

Algoritma tersebut digunakan apabila kata yang dicari sama persis dengan apa yang ada di teks. Dalam hal ini, kata dalam bahasa Jepang merupakan pattern, dan kata dalam bahasa Inggris merupakan teks.

Akan tetapi, kata serapan yang ditulis oleh orang Jepang ke dalam huruf katakana tidak berasal dari bagaimana kata tersebut ditulis dalam bahasa aslinya, namun berasal dari bagaimana kata tersebut diucapkan dan terdengar oleh orang jepang. Seperti yang telah disebutkan sebelumnya, tidak ada huruf konsonan mati pada huruf katakana selain huruf 'n'. Oleh karena itu, kata yang dihasilkan dari huruf katakana sering kali tidak mirip dengan ucapan aslinya.

Jadi, algoritma pattern matching klasik tidak bisa digunakan dalam kasus ini. Untuk kasus pencocokan string katakana ke dalam bahasa Inggris, diperlukan langkah-langkah khusus dan algoritma pattern matching secara aproksimasi.

III. FONEM PATTERN MATCHING

Sebelum dilakukan pentranslasian, kata dalam bahasa Jepang yang ditulis dengan huruf katakana harus diubah terlebih dahulu ke dalam huruf romaji untuk menyeragamkan huruf. Proses pengubahan ini menggunakan *Hepburn romanization system*.

Kemudian kata dalam huruf katakana yang sudah diubah bentuknya menjadi huruf romaji atau alfabet ditranslasikan ke dalam suatu string yang memiliki suatu aturan yang akan menjadikan kata tersebut mirip dengan kata dalam bahasa Inggris.

Untuk mentranslasikan huruf katakana ke dalam bahasa Inggris dan begitu juga sebaliknya, maka diperlukan sebuah tabel translasi yang memuat data-data pentranslasian seperti halnya pada tabel berikut ini.

spelling transformation rule				example	
				katakana word	English origin
u	→	*	/ C_(C|#)	*shisutemu*	system
o	→	*	/ (d|t)_(C|#)	*doraibaa*	driver
i	→	*	/ C_(C|#)	*matchi*	match
howa	→	wh	/ _V	*howaito*	white
(u|uu)	→	w	/ _V	*uuru*	wool
i	→	y		*iesu*	yes
ee	→	yV		*eeru*	Yale
y	→	*	/ _V	*kyaburetaa*	carburetor
a	→	Vr	/ (oo|o|e|i)_#	*hea*	hair
aa	→	Vr		*misutaa*	mister
a	→	Vr	/ _#	*aakitekucha*	architecture
oo	→	Vr		*pooku*	pork
s	→	c	/ _e	*serori*	celery
s	→	th		*sumisu*	Smith
z	→	j	/ _e	*zerii*	jelly
z	→	th		*mazaa*	mother
j	→	(d|z)		*ejison*	Edison
b	→	v		*banira*	vanilla
h	→	f		*ucha*	wafer
r	→	l		*reñgusu*	length
ts	→	(t|z)		*tsurii*	tree

Tabel 1. Data aturan transformasi ejaan

Tabel di atas memiliki aturan sebagai berikut:
C : huruf konsonan
V : huruf vokal
(a|b) : dapat berupa a atau b
* : karakter null
# : batas pinggir kata

Sebaliknya, proses konversi dari bahasa Inggris ke dalam bahasa jepang memiliki masalah. Cara pengucapan dalam bahasa Inggris dapat beragam. Untuk mengatasi hal tersebut, dipilih suatu standar pengucapan yang umum, disebut dengan *IPA phoneme*.

Tabel di bawah ini merupakan tabel *IPA phoneme* yang mentranslasikan pengucapan huruf dalam bahasa Inggris.

English letter	IPA	English letter	IPA
a	/a/	n	/n/
b	/b/	o	/o/
c	/k/	p	/p/
d	/d/	q	/k/
e	/e/	r	/r/
f	/f/	s	/s/
g	/g/	t	/t/
h	/h/	u	/u/
i	/i/	v	/v/
j	/ʤ/	w	/w/
k	/k/	x	/ks/
l	/l/	y	/i/
m	/m/	z	/z/

Tabel 2. *IPA phoneme* huruf dalam bahasa Inggris

Selanjutnya, ditetapkan beberapa aturan dalam cara penulisan dari pengucapan kata dalam bahasa Inggris sesuai dengan seperti apa kata yang terdengar .

no	text-to-phoneme rule	example	
		English word	phoneme sequence
1	ch → (k | ʃ | ʧ)	chemical Chicago bench	/kemikəl/ /ʃikaːgou/ /benʧ/
2	sh → ʃ	show	/ʃou/
3	tion → (ʃon | ʧon)	motion question	/mouʃən /kwesʧən/
4	e → * / C_#	white	/hwait/
5	y → j / C_	yes	/jes/
6	ck → k	pick	/pik/
7	c → s	circle	/səɪrkl/
8	th → (ð| θ)	father length	/faːðər/ /leŋθ/
9	gh → *	neighbor	/neibər/
10	s → z	easy	/iːzi/
11	g → ʤ	energy	/enərʤi/

Tabel 3.*IPA phoneme* kata dalam bahasa Inggris

Kata-kata tersebut nantinya akan dipakai dalam pencocokan kata yang ditulis dengan menggunakan huruf katakana dengan kata yang ditulis dengan menggunakan bahasa Inggris.

Setelah semua bagian sudah disesuaikan masing-masing, tibalah saatnya untuk melakukan pencocokan string dengan menggunakan algoritma.

Algoritma yang digunakan untuk pencocokan string katakana dengan bahasa Inggris adalah sebagai berikut:

```
1 d[0,j]=0, for all j where 0<=j<=n
2 d[i,0]=1, for all i where 0<=i<=m
3 while (0<i<=m and 0<j<=n)
4 {
5    if p[i]=t[j] then
6          A=d[i-1,j-1]
7    else
8          A=d[i-1,j-1]+1
9    B = d[i-1,j]+1
10   C = d[i,j-1]+1
11   d[i,j]=min(A,B,C)
12 }
13  find the smallest value in the
14last row (d[m,j])
```

Penjelasan lebih lanjut terhadap kode di atas adalah sebagai berikut.

Untuk baris ke-1:

```
If char_of_text(j) = space, coma, or period
      then d[0,j]=0
else d[0,j]=d[0,j-1]+1
```

```
d[0,j]=k-j,
where d[0,k] is the entry corresponding to the nearest
delimiter and j<=k
```

Untuk baris ke-5:

```
if t[j] is transformed into the same phoneme p[i] by any
rule in the table base listed in table  then
```

Untuk baris ke-6:

```
if  isvowel(p[i])=TRUE then
      A= d[i-1, j-1]
else
      A = d[i-1,j-1]-1
```

Algoritma tersebut tidak sepenuhnya akurat karena bagaimana pun pasti adanya ketidaksesuaian antara kata bahasa Jepang yang ditulis dengan huruf katakana dengan kata dalam bahasa Inggris. Namun algoritma tersebut jauh lebih baik apabila dibandingkan dengan algoritma pencocokan klasik yang *exact match*.

Algoritma tersebut membutuhkan suatu pentranslasian data untuk menghubungkan antara cara mebaca huruf dalam bahasa Inggris dengan cara membaca katakana yang ditulis dalam romaji. Aturan tersebut mirip dengan aturan yang telah ditetapkan sebelumnya, direpresentasikan dalam tabel berikut ini.

no	phonological rule	example	
		katakana word	English origin
1	u → * / C_(C|#)	/ʃisutemu/	/sistəm/
2	o → * / (d|t)_(C|#)	/doraibaɪ/	/daivər/
3	i → * / C_(C|#)	/maQʧi/	/mæʧ/
4	howa → hw / _V	/howaito/	/hwait/
5	(u|uː) → w / _V	/uːru/	/wul/
6	i → j	/iesu/	/jes/
7	eː → jV	/eːru/	/jeil/
8	j → * / _V	/kyaputeN/	/kæptin/
9	a → Vr / (oː|o|e|i)_#	/hea/	/heər/
10	aː → Vr	/misutaː/	/mistər/
11	a → Vr / _#	/aːkitekuʧa/	/arkiteʧər/
12	oː → Vr	/poːku/	/pɔrk/
13	s → θ	/sumisu/	/smiθ/
14	z → ʤ / _e	/zeriː/	/ʤeli/
15	z → ð	/mazaː/	/maðər/
16	ʤ → (d|z)	/eʤisoN/	/edəsən/
17	b → v	/banira/	/vənilə/
18	h → f	/ueha/	/weifər/
19	r → l	/reNgusu/	/leŋθ/
20	ts → (t|z)	/tsuɪri/	/triː/
21	ʃ → s	/ʃisutemu/	/sistəm/
22	ʧ → t	/maruʧi/	/məlti/
23	oru → Vr / C_	/ʧuːdoru/	/tudər/
24	eru → Vr / C_	/enerugiː/	/enərʤi/

Tabel 4.Aturan fonologi

Setelah dilakukan pentranslasian tersebut, maka akan didapatkan kata dalam bahasa Jepang yang mendekati pengucapan dalam bahasa Inggris.

IV. SIMPULAN

Pada pattern matching dalam penerjemahan kata yang ditulis dengan menggunakan huruf katakana ke dalam bahasa Inggris tidak bisa memakai algoritma klasik yang *exact match*. Pencocokan kata harus melalui beberapa tahap penyesuaian kata berdasarkan fonem terlebih dahulu, kemudian dicocokkan dengan menggunakan algoritma aproksimasi.

V. PUSTAKA

Kang, Yunsun. An Algoritm for Generating Dictionary of Japanese Scientific Terms. Literary and Linguistic Computing, Vol.11, No.2, 1996.

Munir, Rinaldi. *Diktat Kuliah IF3051 Strategi Algoritma*. 2009. Teknik Informatika ITB : Bandung.

http://blogbintang.com/huruf-bahasa-jepang-hiragana-katakana-kanji 1.13
(waktu akses: 9 Desember 2010, pukul 01.10)

http://edukasi.kompasiana.com/2010/06/25/mengenal-huruf-hiragana-katakana-dan-kanji/
(waktu akses: 9 Desember 2010, pukul 01.15)

http://ww3.algorithmdesign.net/handouts/PatternMatching.pdf
(waktu akses: 9 Desember 2010, pukul 04.00)

VI. PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 9 Desember 2010

Nabilah Shabrina (13508087)