9. IBM Model-1 の動作例

内山将夫@NICT
mutiyama@nict.go.jp

小さいコーパス

- 「彼」「の」「絵 」と「his」「painting」
- 「彼」「の」「コレクション」と「his」「collection」
- 「絵」「の」「コレクション」と「painting」「collection」

初期値

f が日本語単語で，eが英単語に相当する．
$t(f|e) = \frac{1}{\text{日本語単語の異なり語数}} = \frac{1}{4}$
$c(f|e) = \Sigma_f \Sigma_s C(f|e; \mathbf{f}^{(s)}, \mathbf{e}^{(s)}) = 0$
と初期値を設定する．

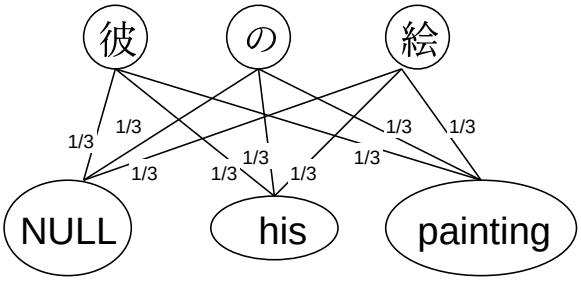
$t(f|e)/c(f|e)$

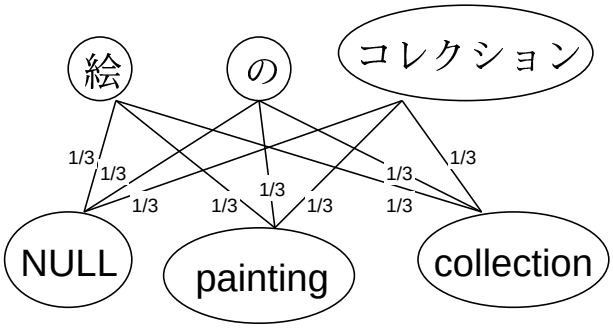
	彼	の	絵	コレクション	確率の計
NULL	$\frac{1}{4}/0$	$\frac{1}{4}/0$	$\frac{1}{4}/0$	$\frac{1}{4}/0$	1
his	$\frac{1}{4}/0$	$\frac{1}{4}/0$	$\frac{1}{4}/0$	$\frac{1}{4}/0$	1
painting	$\frac{1}{4}/0$	$\frac{1}{4}/0$	$\frac{1}{4}/0$	$\frac{1}{4}/0$	1
collection	$\frac{1}{4}/0$	$\frac{1}{4}/0$	$\frac{1}{4}/0$	$\frac{1}{4}/0$	1

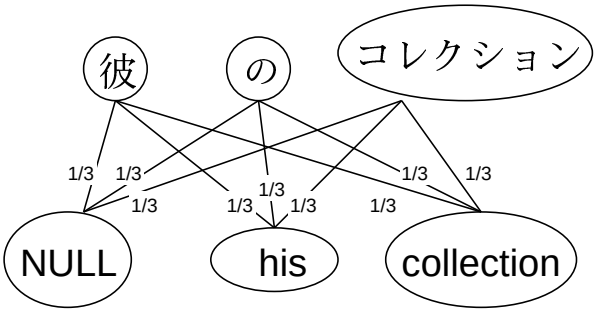
1回目の計算

全てのエッジの重みが$\frac{1}{3}$である．たとえば，
「彼」「の」「絵」と「NULL」「his」「painting」において

$$\frac{t(\text{絵}|\text{painting})}{t(\text{絵}|\text{NULL})+t(\text{絵}|\text{his})+t(\text{絵}|\text{painting})} = \frac{1/4}{1/4+1/4+1/4} = \frac{1}{3}$$







$C(f|e)$の集計 / $t(f|e)$の再推定

$$C(f|e) = f\text{と}e\text{をつなぐエッジの重みの総和}$$

$$t(f|e) = \frac{C(f|e)}{\Sigma_f C(f|e)} = \frac{C(f|e)}{C(e)}$$

「NULL」と「彼」に注目すると，このペアは2回でたので

$$C(\text{彼}|\text{NULL}) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

一方

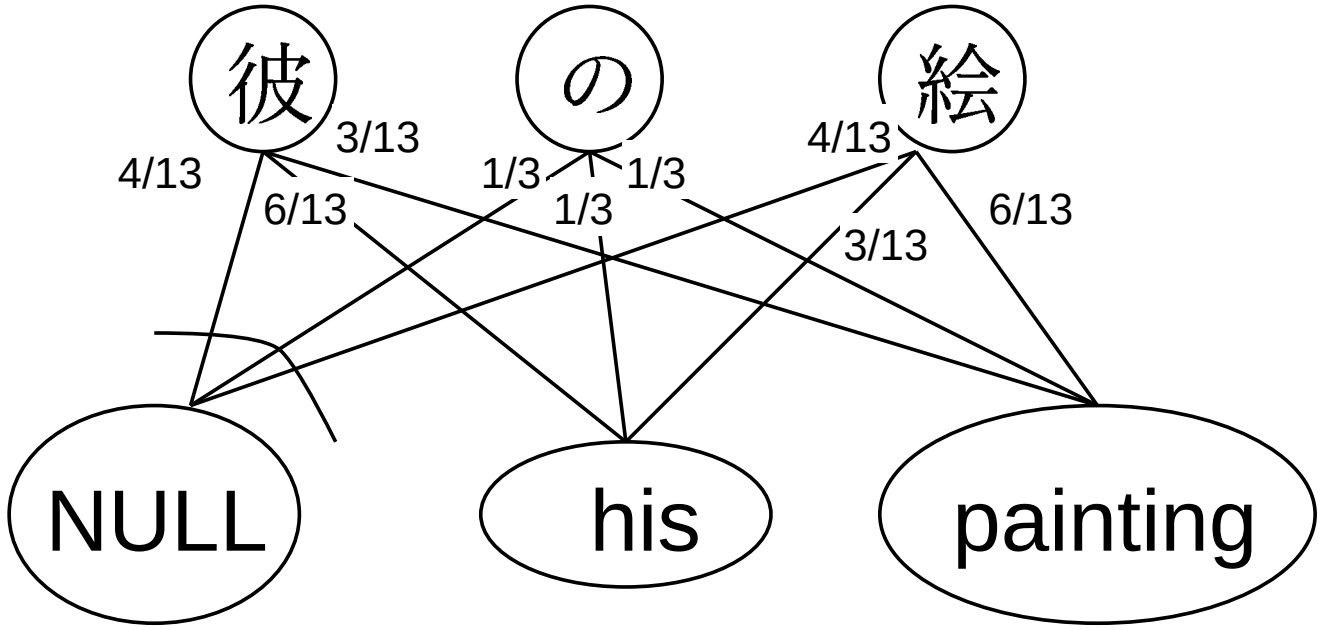
$$C(\text{NULL}) = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} + \frac{1}{3} + \frac{1}{3} + \frac{1}{3} + \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = \frac{9}{3}$$

よって，

$$t(\text{彼}|\text{NULL}) = \frac{C(\text{彼}|\text{NULL})}{C(\text{NULL})} = \frac{\frac{2}{3}}{\frac{9}{3}} = \frac{2}{9}$$

	彼	の	絵	コレ	$C(e)$
NULL	$\frac{1}{3} + \frac{1}{3}/\frac{2}{9}$	$\frac{1}{3} + \frac{1}{3} + \frac{1}{3}/\frac{3}{9}$	$\frac{1}{3} + \frac{1}{3}/\frac{2}{9}$	$\frac{1}{3} + \frac{1}{3}/\frac{2}{9}$	$\frac{9}{3}$
his	$\frac{1}{3} + \frac{1}{3}/\frac{2}{6}$	$\frac{1}{3} + \frac{1}{3}/\frac{2}{6}$	$\frac{1}{3}/\frac{1}{6}$	$\frac{1}{3}/\frac{1}{6}$	$\frac{6}{3}$
paint.	$\frac{1}{3}/\frac{1}{6}$	$\frac{1}{3} + \frac{1}{3}/\frac{2}{6}$	$\frac{1}{3} + \frac{1}{3}/\frac{2}{6}$	$\frac{1}{3}/\frac{1}{6}$	$\frac{6}{3}$
coll.	$\frac{1}{3}/\frac{1}{6}$	$\frac{1}{3} + \frac{1}{3}/\frac{2}{6}$	$\frac{1}{3}/\frac{1}{6}$	$\frac{1}{3} + \frac{1}{3}/\frac{2}{6}$	$\frac{6}{3}$

2回目の計算：文1

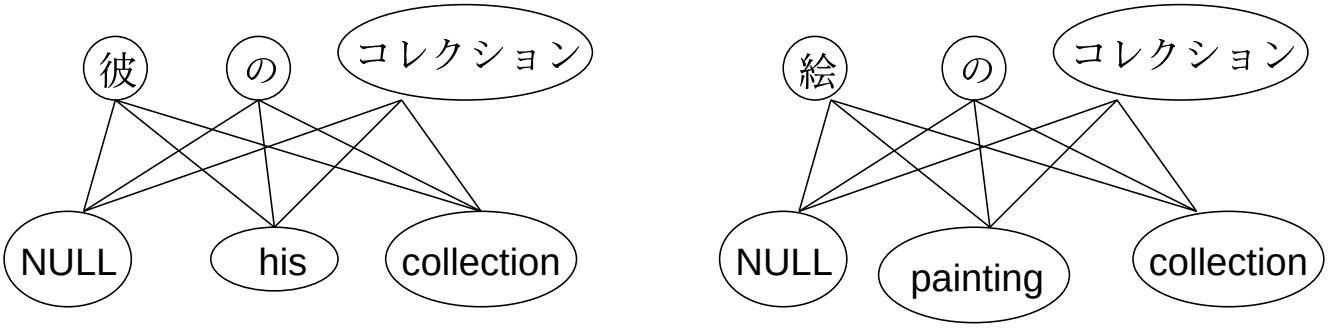


$$\frac{t(\text{彼}|\text{NULL})}{t(\text{彼}|\text{NULL})+t(\text{彼}|\text{his})+t(\text{彼}|\text{painting})} = \frac{\frac{2}{9}}{\frac{2}{9}+\frac{2}{6}+\frac{1}{6}} = \frac{4}{13}$$

$$\frac{t(\text{の}|\text{NULL})}{t(\text{の}|\text{NULL})+t(\text{の}|\text{his})+t(\text{の}|\text{painting})} = \frac{\frac{3}{9}}{\frac{3}{9}+\frac{2}{6}+\frac{2}{6}} = \frac{1}{3}$$

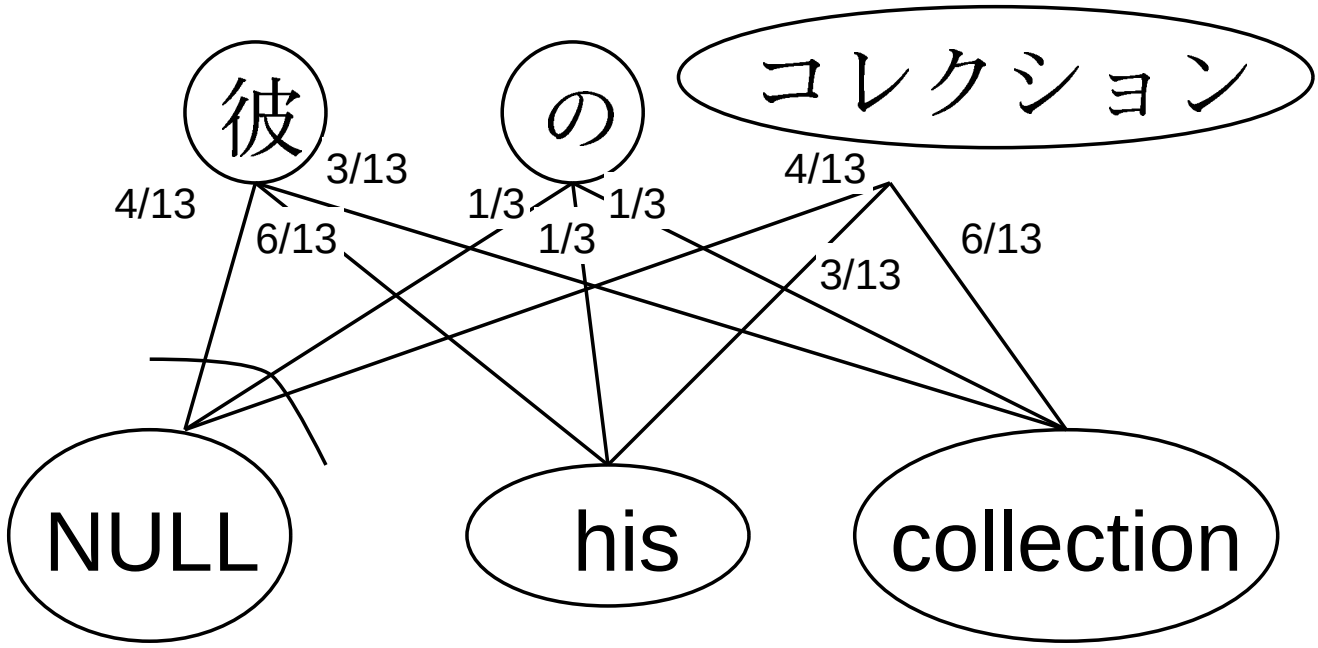
$$\frac{t(\text{絵}|\text{NULL})}{t(\text{絵}|\text{NULL})+t(\text{絵}|\text{his})+t(\text{絵}|\text{painting})} = \frac{\frac{2}{9}}{\frac{2}{9}+\frac{1}{6}+\frac{2}{6}} = \frac{4}{13}$$

問題 (15分)



上記2文における各エッジの重みを求めること

2回目の計算：文2

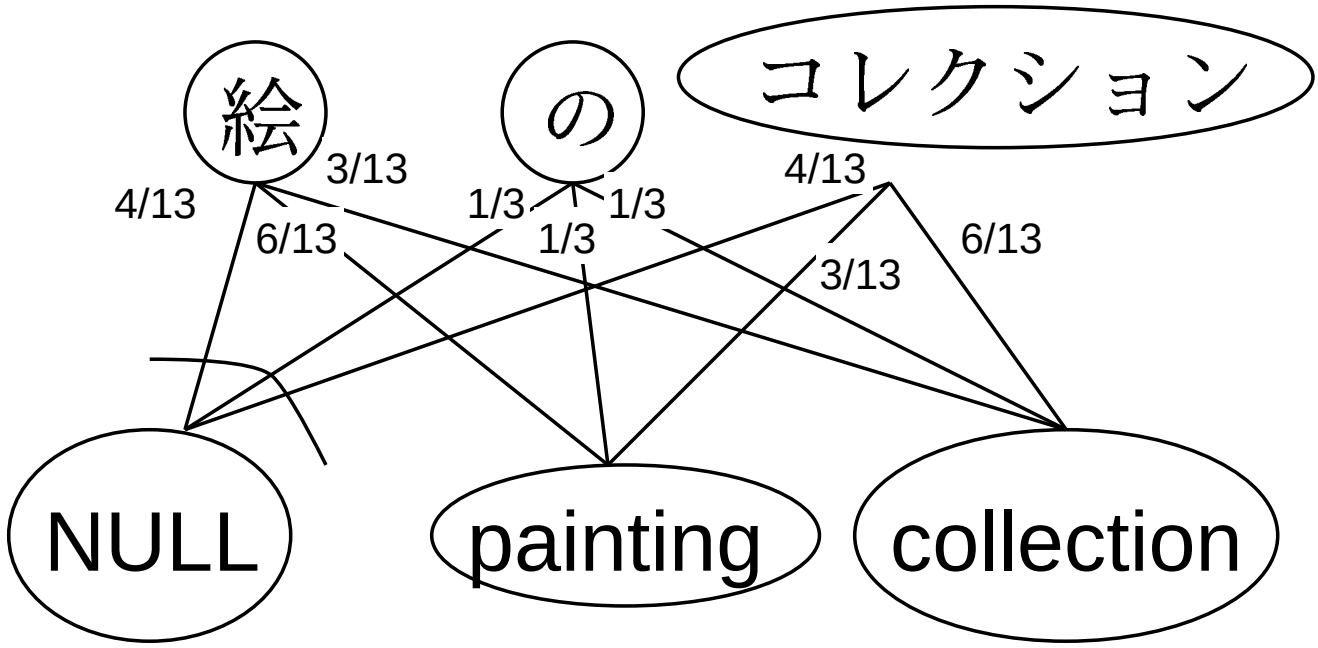


$$\frac{t(\text{彼}|\text{NULL})}{t(\text{彼}|\text{NULL})+t(\text{彼}|\text{his})+t(\text{彼}|\text{collection})} = \frac{\frac{2}{9}}{\frac{2}{9}+\frac{2}{6}+\frac{1}{6}} = \frac{4}{13}$$

$$\frac{t(\text{の}|\text{NULL})}{t(\text{の}|\text{NULL})+t(\text{の}|\text{his})+t(\text{の}|\text{collection})} = \frac{\frac{3}{9}}{\frac{3}{9}+\frac{2}{6}+\frac{2}{6}} = \frac{1}{3}$$

$$\frac{t(\text{コレクション}|\text{NULL})}{t(\text{コ}|\text{NULL})+t(\text{コ}|\text{his})+t(\text{コ}|\text{collection})} = \frac{\frac{2}{9}}{\frac{2}{9}+\frac{1}{6}+\frac{2}{6}} = \frac{4}{13}$$

2回目の計算：文3



$$\frac{t(\text{絵}|\text{NULL})}{t(\text{絵}|\text{NULL})+t(\text{絵}|\text{painting})+t(\text{絵}|\text{collection})} = \frac{\frac{2}{9}}{\frac{2}{9}+\frac{2}{6}+\frac{1}{6}} = \frac{4}{13}$$

$$\frac{t(\text{の}|\text{NULL})}{t(\text{の}|\text{NULL})+t(\text{の}|\text{painting})+t(\text{の}|\text{collection})} = \frac{\frac{3}{9}}{\frac{3}{9}+\frac{2}{6}+\frac{2}{6}} = \frac{1}{3}$$

$$\frac{t(\text{コレクション}|\text{NULL})}{t(\text{コ}|\text{NULL})+t(\text{コ}|\text{painting})+t(\text{コ}|\text{collection})} = \frac{\frac{2}{9}}{\frac{2}{9}+\frac{1}{6}+\frac{2}{6}} = \frac{4}{13}$$

問題 (15分)

前述の$C(f|e)$の集計 / $t(f|e)$の再推定における

	彼	の	絵	コレ	$C(e)$
NULL	$\frac{1}{3}+\frac{1}{3}/\frac{2}{9}$	$\frac{1}{3}+\frac{1}{3}+\frac{1}{3}/\frac{3}{9}$	$\frac{1}{3}+\frac{1}{3}/\frac{2}{9}$	$\frac{1}{3}+\frac{1}{3}/\frac{2}{9}$	$\frac{9}{3}$
his	$\frac{1}{3}+\frac{1}{3}/\frac{2}{6}$	$\frac{1}{3}+\frac{1}{3}/\frac{2}{6}$	$\frac{1}{3}/\frac{1}{6}$	$\frac{1}{3}/\frac{1}{6}$	$\frac{6}{3}$
paint.	$\frac{1}{3}/\frac{1}{6}$	$\frac{1}{3}+\frac{1}{3}/\frac{2}{6}$	$\frac{1}{3}+\frac{1}{3}/\frac{2}{6}$	$\frac{1}{3}/\frac{1}{6}$	$\frac{6}{3}$
coll.	$\frac{1}{3}/\frac{1}{6}$	$\frac{1}{3}+\frac{1}{3}/\frac{2}{6}$	$\frac{1}{3}/\frac{1}{6}$	$\frac{1}{3}+\frac{1}{3}/\frac{2}{6}$	$\frac{6}{3}$

の表を更新すること

$C(f|e)$の集計 / $t(f|e)$の再推定

$$C(f|e) = f\text{と}e\text{をつなぐエッジの重みの総和}$$

$$t(f|e) = \frac{C(f|e)}{\Sigma_f C(f|e)} = \frac{C(f|e)}{C(e)}$$

「NULL」と「彼」に注目すると，

$$C(\text{彼}|\text{NULL}) = \frac{4}{13} + \frac{4}{13} = \frac{8}{13}$$

$$C(\text{NULL}) = \frac{4}{13} + \frac{4}{13} + \frac{1}{3} + \frac{1}{3} + \frac{1}{3} + \frac{4}{13} + \frac{4}{13} + \frac{4}{13} + \frac{4}{13} = \frac{37}{13}$$

よって，

$$t(\text{彼}|\text{NULL}) = \frac{C(\text{彼}|\text{NULL})}{C(\text{NULL})} = \frac{\frac{8}{13}}{\frac{37}{13}} = \frac{8}{37}$$

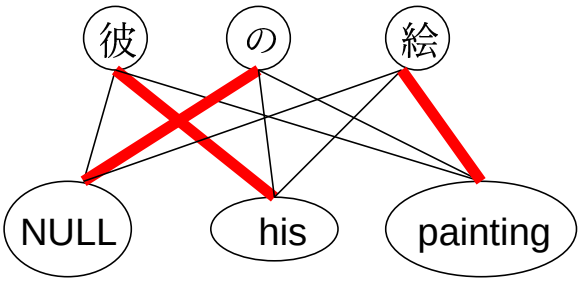
	彼	の	絵	コレ	$C(e)$
NULL	$\frac{4}{13} + \frac{4}{13}/\frac{8}{37}$	$\frac{1}{3} + \frac{1}{3} + \frac{1}{3}/\frac{13}{37}$	$\frac{4}{13} + \frac{4}{13}/\frac{8}{37}$	$\frac{4}{13} + \frac{4}{13}/\frac{8}{37}$	$\frac{37}{13}$
his	$\frac{6}{13} + \frac{6}{13}/\frac{36}{80}$	$\frac{1}{3} + \frac{1}{3}/\frac{26}{80}$	$\frac{3}{13}/\frac{9}{80}$	$\frac{3}{13}/\frac{9}{80}$	$\frac{80}{39}$
paint.	$\frac{3}{13}/\frac{9}{80}$	$\frac{1}{3} + \frac{1}{3}/\frac{26}{80}$	$\frac{6}{13} + \frac{6}{13}/\frac{36}{80}$	$\frac{3}{13}/\frac{9}{80}$	$\frac{80}{39}$
coll.	$\frac{3}{13}/\frac{9}{80}$	$\frac{1}{3} + \frac{1}{3}/\frac{26}{80}$	$\frac{3}{13}/\frac{9}{80}$	$\frac{6}{13} + \frac{6}{13}/\frac{36}{80}$	$\frac{80}{39}$

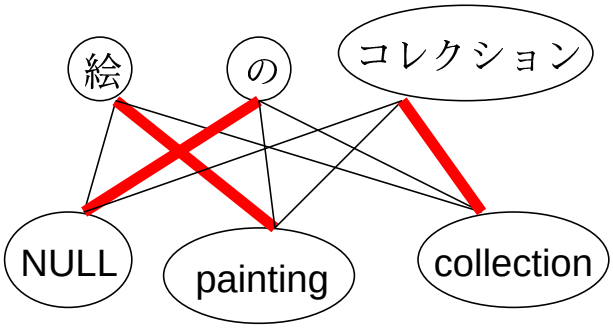
赤いセルに確率が集中していくことがわかる．

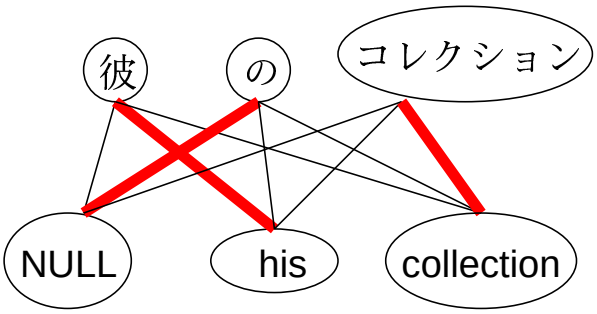
	彼	の	絵	コレ	$C(e)$
NULL	.615/.216	1.0/.351	.615/.216	.615/.216	2.846
his	.923/.45	.667/.325	.231/.113	.231/.113	2.051
paint.	.231/.113	.667/.325	.923/.45	.231/.113	2.051
coll.	.231/.113	.667/.325	.231/.113	.923/.45	2.051

単語アラインメント

確率最大の$t(f|e)$について赤くする．







日英対訳文対応付けデータ(内山・高橋2003)での例

約10万文の小説等のコーパスからIBM Model-1で得られた対訳確率が0.5以上でかつ無作為に抽出した単語対の例

英日方向

経済/economic, ワルツ/waltz, 宮殿/palace, 音楽/music, フリー/free, ウワア/wow, ：/:, 1999/1999, US/us, ファーディナンド/ferdinand, 〃/., 百万/million, 砂地/sand, ７．/7., ほこり/dust, 伯父/uncle, 霜/frost, ブリ/gabriel, 停留所/stop, ポール/paul, ウマ/horse, オックスフォード/oxford, 両方/both, サー/sir, 財産/property, ハドソン/hudson, 薄い/thin, エウリュピュロス/eurypylus, ベルリン/berlin, ２０３．/the, ?/?, 高等/higher, 音節/syllables, イグネイシャス/ignatius, ピナー/pinner, 日/day, ヶ月/months, １０１．/the, 銅/copper, コンロイ/conroy, ハウス/house, ベーアマン/behrman, ほぼ/almost, 唇/lips, 新/new, 主題/subject, 祭壇/altar, エドワード/edward, 太陽/sun, Software/software, 現象/phenomena, 空中/air, 友情/friendship, ガ/gallaher, 戦う/fight, 君たち/you, 気付/hearthrug, 三月/march, 被曝/exposure, ５９．/mischievous, ヘンリー/henry, ワトスン/watson, マシン/machine, イタリア/italy,)/), エンジニアリング/engineering, 1986/1986, ２２９．/the, 美しい/beautiful, ドードー/dodo, 損失/loss, 呑み/gasped, シカゴ/chicago, Web/web, 砂漠/desert, 地主/squire, 朝食/breakfast, バーベキュー/barbecue, 1509/1509, デザイナー/designer, 下っ/down, アクセス/access, Law/law, 青/blue, 星/stars, ジョンストン/johnston, 蓮/lotus, ズン/thump, アンリエッタ/henrietta, エルサレム/jerusalem, 風の音/wind

日英方向

nose/鼻, amen/アーメン, ..7-1/., network/ネットワーク, dorothy/ドロシー, .ix/., joe/ジョー, winter/冬, animals/動物, perrault/ペロー, window/窓, julia/ジュリア, church/教会, smoke/煙, reasons/理由, trousers/ズボン, patten/パッテン, armour/武具, tooth/歯, passepartout/パスパルトゥー, troy/トロイア, es/., era/時代, freedom/自由, flag/旗, dublin/ダブリン, churches/教会, table/テーブル, eagle/ワシ, darwin/ダーウィン, poem/詩, daisy/デイジー, daughter/娘, shoulders/肩, twyford/トワイフォード, 2/2, branches/枝, maimie/マイミー, “/「, usecbc/UseCBC, master/主人, jim/ジム, uh/はい, purposes/目的, 20/20, straits/海峡, importation/輸入, endfor/EndFor, film/映画, fiddle-stick/ばかばかしい, 1997/1997, fox/キツネ, history/歴史, 12/12, flower/花, german/ドイツ, years/年, eveline/エヴリン, religion/宗教, register/レジスタ, skirt/スカート, poole/プール, lisp/LISP, slowly/ゆっくり, tink/ティンク, mit/MIT, kiotsukete/キヲツケテ, woods/森, puck/パック, head/頭, lysander/ライサンダー, french/フランス, insurance/保険, reproduction/複製, ..3-1/., horses/馬, ..5-1/., .41/., israeli/イスラエル, christianity/キリスト教, priam/プリアモス, ohio/オハイオ, intellect/知性, shoulder/肩, wall/壁, buildings/建物, truths/真理, rights/権利, gnu/GNU, gentleman/紳士, 16/16, program/プログラム

アラインメントの例

X: 日英方向
Y: 英日方向
両方向で選ばれた対応：

“と「，angelと天使，likeとよう

	「	天使	の	よう	だ	ね	」
“	XY				X		
he							
looks		Y				X	
just		Y					
like				XY			
an		Y	X				
angle		XY					
“							Y

まとめ

- IBM Model-1 の実行例を観察した
- 対訳単語の抽出や単語対応がとれることをみた
- Model-1はそれほど強力なモデルではない

課題

興味があれば，`http://www2.nict.go.jp/x/x161/members/mutiyama/software.html`にあるIBM Model-1 のプログラムを実行してみる