15. フレーズテーブルの作り方

内山将夫@NICT
mutiyama@nict.go.jp

フレーズテーブルの作成手順

1. 初期単語アラインメント
2. 単語アラインメントの修正
3. 単語単位の翻訳確率の推定
4. フレーズ対応の抽出
5. フレーズへのスコアの付与

結果

英語句 ||| 日本語句 ||| スコア1 スコア2 ...

これらの句のスコアを上手く組み合わせて文全体のスコアを得る.

初期単語アラインメント

GIZA++ というフリーソフトウェアを利用することが多い．これを，たとえば，日英の方向に，

- IBM Model-1 (紹介済)
- HMM Model
- IBM Model-3
- IBM Model-4

を順番に適用していくことにより，日英の各文対応について，1対nの単語対応を得ることができる．
英日にもすることにより，両方向に，1対nの単語対応を得る．

==========

単語アラインメントは未解決の問題だが，ある程度まではできるので，とりあえず，そこまでを前提にして，次に進んでいる．

日英単語アラインメントの例

	when	the	fluid	pressure	cylinder	31	is	used	,	fluid	is	gradually	applied	.
流体			*											
圧				*										
シリンダ					*									
３１						*								
の														
場合	*													
は							*							
流体										*				
が											*			
徐々に												*		
排出												*		
さ													*	
れる													*	
こと													*	
と									*					
なる														
。														*

日英単語アラインメントの例

	this	relation	must	be	maintained	even	after	passing	at	least	100,000	sheets	.
そして							*						
、				*									
上記	*												
関係		*											
を								*					
少なくとも										*			
１０万											*		
枚											*		
通											*		
紙												*	
し								*					
て						*							
も						*							
維持					*								
し								*					
なけれ			*										
ば			*										
なら			*										
ない			*										
。													*

日英単語アラインメントの例

	first	,	a	description	will	be	given	of	the	structure	of	the	wheels	2	.
まず	*														
、		*													
車輪													*		
2														*	
の											*				
構造										*					
について					*										
説明					*										
する					*										
。															*

英日単語アラインメントの例

	流体	圧	シリンダ	３１	の	場合	は	流体	が	徐々に	排出	さ	れる	こと	と	なる
when						*										
the							*									
fluid	*															
pressure		*														
cylinder			*													
31				*												
is													*			
used															*	
,																
fluid								*								
is									*							
gradually										*						
applied											*					
.																

英日単語アラインメントの例

	まず	、	車輪	2	の	構造	について	説明	する	。
first	*									
,		*								
a										
description								*		
will								*		
be								*		
given								*		
of					*					
the					*					
structure						*				
of							*			
the									*	
wheels			*							
2				*						
.										*

日英英日アラインメントの合成

- アラインメントの積集合 = I
  → 精度の高い単語対応
- アラインメントの積集合 = U
  → カバー率の高い単語対応

I を出発点として U 中のアラインメントを加えていく.

日英と英日の合成の例

	when	the	fluid	pressure	cylinder	31	is	used	,	fluid	is	gradually	applied	.
流体			I											
圧				I										
シリンダ					I									
３１						I								
の														
場合	I													
は		U					U							
流体										I				
が											I			
徐々に												I		
排出												U	U	
さ													U	
れる							U						U	
こと													U	
と								U	U					
なる														
。														I

日英と英日の合成の例

	this	relation	must	be	maintained	even	after	passing	at	least	100,000	sheets	.
そして							U						
、				U									
上記	I												
関係		I											
を								U					
少なくとも									U	I			
１０万											I		
枚											U	U	
通								U			U		
紙												U	
し							U	U					
て						U							
も						I							
維持					I								
し								U					
なけれ			U										
ば			U										
なら			I	U									
ない			U										
。													I

日英と英日の合成の例

	first	,	a	description	will	be	given	of	the	structure	of	the	wheels	2	.
まず	I														
、		I													
車輪													I		
2														I	
の								U	U		U				
構造										I					
について					U						U				
説明				U	I	U	U								
する					U							U			
。															I

合成手順

```
#
# アラインメント候補 = {(e_i, f_j)} は e_i + fj が小さいものから並
# んでいる．そうすることにより，文頭のものから優先して対応付けてい
# くことができる．
#

日英英日アラインメントの合成 ()
  隣接点集合 = ((-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1))
  アラインメント候補 = 積集合 I(日英アラインメント，英日アラインメント)
  隣接する 1 対 1 の点を追加する ()
  その他の点を加える ()
  アラインメント候補を返す
end

隣接する 1 対 1 の点 N を追加する ()
  while true
    for (j, e) ∈ アラインメント候補
      for (j0, e0) ∈ 隣接点 (j,e) ∩ 和集合 (日英，英日)
         if(j0 も e0 もアラインメント候補で使われていない)
           (j0, e0) をアラインメント候補に追加する (そうしても 1 対 1 が壊れない)
         end
      end
    end
    アラインメント候補に追加がなければ終了する
  end
end

#
# なるべく多くの点を加えたいが，日英共に使われているのは加えない
#
その他の点 O を加える ()
  for (j,e) ∈ 和集合 (日英，英日)
    if (j か e のどちらかがアラインメント候補で使われていない)
      (j,e) をアラインメント候補に含める
    end
  end
end
```

合成の例

(I=積集合，N=隣接点，O=その他の追加, U=削除された点)

	when	the	fluid	pressure	cylinder	31	is	used	,	fluid	is	gradually	applied	.
流体			I											
圧				I										
シリンダ					I									
31						I								
の														
場合	I													
は		N					O							
流体										I				
が											I			
徐々に												I		
排出												U	N	
さ													O	
れる							U						O	
こと													O	
と								O	O					
なる														
。														I

合成の例

	this	relation	must	be	maintained	even	after	passing	at	least	100,000	sheets	.
そして							O						
、				O									
上記	I												
関係		I											
を								O					
少なくとも									O	I			
１０万											I		
枚											U	N	
通								U			O		
紙												O	
し							U	O					
て						O							
も						I							
維持					I								
し								O					
なけれ			O										
ば			O										
なら			I	U									
ない			O										
。													I

合成の例

	first	,	a	description	will	be	given	of	the	structure	of	the	wheels	2	.
まず	I														
、		I													
車輪													I		
2														I	
の								O	N		U				
構造										I					
について					U						N				
説明				O	I	O	O								
する					O							O			
。															I

単語単位の翻訳確率

- これまでに，日英，英日の1対nおよびm対1の単語対応を組み合わせてm対nの単語対応を得る方法を示した.
- これから，それを利用して，フレーズを抽出する方法を示す.
- その前に，ここで得たm対nの対応から単語単位の辞書を作る方法を示す.

単語単位の辞書

アライメント

	e1	e2	e3	e4
j1	*		*	
j2				
j3		*		
j4				*

から，(j1,e1), (j1,e3), (j3, e2), (j4,e4) の単語対応がとれるので，これから，j $\rightarrow$ e と e $\rightarrow$ j 方向の辞書がつくれる．更に，

$$P(j|e) = \frac{\#(j,e)}{\Sigma_{j'}\#(j',e)}$$

$$P(e|j) = \frac{\#(j,e)}{\Sigma_{e'}\#(j,e')}$$

により，単語単位の翻訳確率が求まる．(スムージングをしても良い)

単語単位の辞書：頻度 日本語単語 英語単語

2230457 、 the
1811473 、 ,
1804260 。 .
1802874 の of
1767572 の the
704450 に to
668066 は the
584253 が is
543500 は is
527825 に in
480009 する the
448118 を is
376711 と and
376195 、 a
365107 図 fig.
345039 を of
287703 ） ）
285186 （ （
281659 信号 signal
272811 を to
245721 データ data
230038 回路 circuit
224630 で in
209459 で ,
202160 に on
198721 １ 1
196317 から from
193708 この the
177043 に shown
175086 し the
166888 、 and
164284 た the
154881 は a
146977 さ to
146509 ２ 2
143482 を for
142192 し to
140494 出力 output
139684 制御 control
135600 が are
130609 は are
128437 この this
126068 実施 embodiment
124007 ３ 3
117825 する to
112598 よう as
111611 た a
110575 ステップ step
109286 ４ 4
106286 および and
105113 を are
101722 で by
101291 及び and
100440 示す shown
98899 が of
98684 が can
97290 示す in
96570 ５ 5
95364 １０ 10
95330 電圧 voltage
93903 する be
93619 入力 input
92337 さ in
89141 形成 formed
87623 する a
87410 において in
85936 １ first
85471 その the
84919 として as
84888 第 first
84615 に with
84356 メモリ memory
82459 ６ 6
82098 値 value
81423 に ,
79752 情
報 information
79287 １２ 12
79126 １１ 11
79048 の for
78406 例 embodiment
78238 に at
78233 こと can
77678 接続 connected
77671 、 an
77513 部 portion
75575 画像 image
75000 ／ /
73687 に into
72920 上記 the
72590 層 layer
72466 第 second
72399 ７ 7
72231 膜 film

単語単位の辞書：$P(日|英)$でソート

emu emu
SCLK sclk1
+ .sub.+
ウェル -well
LRD lrd
LAT lat.sub.
MRSOUT mrsout
0 0 0 0 0 0 0 0 00000000
WDR wdr
CBS cbs.sub.
PxR pxr
MBT mbt
MBQ mbq
BRE bre
685 685
. 38q
ramarea ramarea
kDa kda
Sdw s.sub.dw
RHIT rhit
CS cs.sub.
dspls dspls
ZODL zodl
GCF gcf
7030 7030
pretrn pretrn
dqmx dqmx
2 2.sup.
wreq wreq
RWC rwc
SGP sgp
3080 3080
排気 exhaling
カテプシン cathepsin
WCBRB wcbrb
Vpk vpk
Vfv vfv
TRBL trbl
PDRHIT pdrhit
Cts c.sub.ts
4080 4080
試験 under-test
romarea romarea
RMIS rmis
PG pg.sub.
CTRLR ctrlr
311 3.sub.11
3090 3090
2041 204.sub.1
10070 10070
セルフプリチャージ self-precharge
waitid waitid
ZMR zmr
Vpass vpass
OPRH oprh
LIOi lioi
1916 1916
日 date-indicating
rblj rblj
ZCO zco
WSEL wsel.sub.
VTTAO vttao
RSRav rsr.sub.av
PDRTPi pdrtpi
Nsa nsa.sub.
Iab i.sub.ab
Cdx cd.sub.x
× .times..sigma..sub.i3
inode inode
SENPn senpn
SDBn sdbn
PFAi pfai
MOPLS mopls
K F L E A K M X kfleakmx
CAD cad7
980 980.degree.
8510 8510
831 83.sub.1
10 ta.sub.10
偏り one-sidedness
微動 micromotion
q se.sub.1-q
ment ment
e E C O N O M Y eeconomy
TRRQ trrq
RSDF rsdf
RPRG rprg
QHN qhn
IRQ irq9
HNMOS hnmos
GLj glj
EAV eav
DWDE dwde
Bjb bjb
BRIa bria
5022 5022
3122 3122

単語単位の辞書：$P(英|日)$でソート

オーディオ audio
ｐｃｈＭＯＳＦＥＴ pch-mosfet
ＩＩＤＲ iidr
@ @
ｃｏａｒｓｅ coarse
ｉｒｏｍ irom
Ｖｃｃｐ vccp
ＧＤＩ gdi
ＭＳＫｉ mski
ＣＫＭ ckm
ＮＰＲ npr
ＳＰＬ spl
ＲＨＩＴ rhit
ＮＯＸ .sigma.nox
ＺＭＣＨＧ zmchg
ＳＴＸ stx
ＲＩＴ rit
ＧＢＵＦＲ gbufr
ｅｘｔＺＲＡＳ extzras
ＭＮＰ mnp
コースター coaster
Ｓｕｐ s.sub.up
９１００ 9100
ｅｘｔＺＣＡＳ extzcas
ＳＨＬ shl
Ｉｎｚ in.sub.z
Ｉｍｓ i.sub.ms
ＢＬｒ blr
フォーマツタ formater
ｄｉｇｅｓｔ digest
ＲＸＴｐｕｌｓｅ rxtpulse
ＤＲａ dra
ＷＢＬｎ wbln
ＳＰＸ spx
ＳＥＬＦＯＳＣ selfosc
ＦＭＴ fmt
ＣＡＳＩ casi
ＰＲＥＤ pred
ＤＴＧ dtg
Ｃｏｐｙ copy
ＣＡａ caa
ｔＰＧＲ tpgr
ｄｑｍｘ dqmx
ＸＳＵ xsu
Ｖｒｃ vrc
ＶｃｏｍＨ vcomh
Ｓｌｕ slu
ＤＢＢａ dbba
ｍｐｇｍ mpgm
ｆＡ fa
ＸＢＬＡ xbla
ＲＥＭＥ reme
ＭＢＷ mbw
Ｃｌｔ clt
ＣＯＣ coc
ＣＬＥ cle
ＲＷＬＥ rwle
ＲＢＬｎ rbln
ＮＢＣ nbc
ＢＴＧ btg
ＢＧＭＭ bgmm
８２，８２ 82
スレッシュ threshold
ｑｃ qc
ｏａｍ oam
ｍｅｔａｐｈｏｒ metaphor
ＴＨｉｄｌ th.sub.idl
ＳＰＫＵ spku
ＲＭＢ rmb
ＲＦＩＣ rfic
ＥＸＴＣＬＫ extclk
ＢＬｈ blh
４０９０ 4090
ｎＭＩＳ nmis
ｌｕｎｃｈｅｒ luncher
ａｕｔｏｍａｔｏｎ automaton
Ｖｈｉｇｈ vhigh
ＳＬＩＦ slif
ＧＷＤ gwd
ＧＦＭ gfm
ＥＡＤ ead
ＤｉｎＬＢｎ dinlbn
ＣＷＬ cwl
ＣＴＲＬＲ ctrlr
ＢＳＢ bsb
１，５５５ 1,555
ｗａｉｔｉｄ waitid
ｖｂａｔ vbat
ｆｕｐ f.sub.up
ＸＳＤ xsd
ＶＺ vz
ＳＡＮＧ sang
ＭＩＳＦＥＴＱｈ q.sub.h

単語単位の辞書：対数尤度比でソート

。 .
の of
、 ,
、 the
の the
図 fig.
信号 signal
）)
（ (
データ data
に to
が is
回路 circuit
と and
は is
１ 1
から from
に in
出力 output
制御 control
２ 2
３ 3
を is
４ 4
電圧 voltage
実施 embodiment
１０ 10
ステップ step
５ 5
入力 input
メモリ memory
情報 information
１２ 12
６ 6
１１ 11
値 value
この this
は the
層 layer
発明 invention
７ 7
／ /
画像 image
膜 film
８ 8
方向 direction
２０ 20
１４ 14
に on
および and
する the
１３ 13
処理 processing
アドレス address
位置 position
、 a
形成 formed
及び and
で in
ゲート gate
９ 9
１６ 16
接続 connected
第 first
よう as
電流 current
電極 electrode
２１ 21
１５ 15
本 present
動作 operation
２２ 22
部 portion
基板 substrate
半導体 semiconductor
として as
図 figs.
３０ 30
示す shown
トランジスタ transistor
例 embodiment
に shown
第 second
レベル level
端子 terminal
装置 apparatus
装置 device
を of
状態 state
記録 recording
１ first
ない not

$P(*|\text{流体})$ と $P(*|\mathbf{fluid})$

```
# P(*|流体)
流体 fluid 0.84868421
流体 hydraulic 0.01946272
流体 gas 0.01151316
流体 fluids 0.01096491
流体 flow 0.01041667
流体 liquid 0.00849781
流体 liquid-pressure 0.00603070
流体 hydrodynamic 0.00575658
流体 working 0.00301535
流体 piston 0.00274123
```

```
# P(*|fluid)
fluid 流体 0.45072063
fluid 液 0.23875382
fluid 油 0.10642015
fluid 液体 0.02576794
fluid ブレーキフルード 0.01892561
fluid 流動 0.01368467
fluid 油圧 0.00873490
fluid 流 0.00844373
fluid 作動 0.00698792
fluid 連 0.00684234
```

フレーズ対応の抽出

これまでに

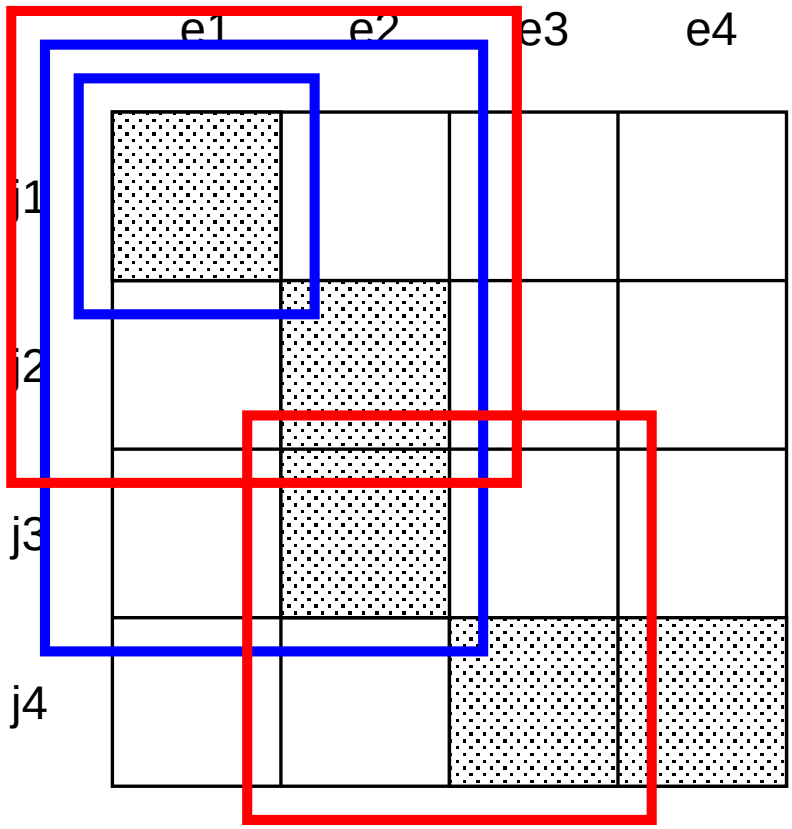
	e1	e2	e3	e4
j1	*			
j2		*		
j3		*		
j4			*	*

のような単語対応を得る方法を述べた．次は，ここから句対応を得る方法を述べる．

基本方針

単語対応に整合するような，なるべく多くの句対応を得る．

整合的な句の例



青はＯＫ．赤はダメ．
日本語の句Jと英語の句Eについて，$j \in J$の対応先を$e(j)$とすると，$e(j) \in E$でないといけない．
同様に$e \in E$の対応先を$j(e)$とすると，$j(e) \in J$である．
つまり，対応の相手は，日本語句Ｊ内の単語と英語句Ｅ内の単語に限られる．

句対応の例：fluid を含む句の例

(流体 ||| (fluid ||| 0.166667 0.0307503 1 0.623726 2.718
(流体 圧) ||| (fluid pressure) ||| 0.5 0.000742744 1 0.200361 2.718
(流体 圧 ||| (fluid pressure ||| 0.5 0.010301 1 0.266855 2.718
(流体 圧 源) ||| (fluid pressure source) ||| 1 0.000104349 1 0.159524 2.718
(流体 圧 源) １７ ||| (fluid pressure source) 17 ||| 1 9.465e-05 1 0.132599 2.718
(流体 圧 源 ||| (fluid pressure source ||| 1 0.00144721 1 0.212466 2.718
) の 送 液 |||) of fluid ||| 1 2.88816e-05 1 0.0150938 2.718
) の 送 液 が |||) of fluid has ||| 1 6.6955e-06 1 0.000451654 2.718
) の 送 液 が あっ た |||) of fluid has taken ||| 1 2.55555e-09 1 7.43761e-07 2.718
、 ３ は 液 圧制 御 装置 ||| , a fluid pressure control apparatus 3 ||| 1 2.28706e-07 1 0.000292281 2.7
、 ３ は 加工 液 ||| , 3 a dielectric fluid ||| 1 3.63074e-05 1 3.96712e-06 2.718
、 ３ は 加工 液 、 ||| , 3 a dielectric fluid , and ||| 1 1.47644e-05 1 5.67333e-08 2.718
、 ３８ は 加工 液 ||| , 38 is working fluid ||| 1 0.0088037 1 0.000662696 2.718
、 ３８ は 加工 液 供給 ||| , 38 is working fluid supplying ||| 1 0.0062226 1 4.34916e-05 2.718
、 ３８ は 加工 液 供給 手段 ||| , 38 is working fluid supplying means ||| 1 0.00427184 1 3.16053e-05 2
、 ４１ ｄ 、 流体 ||| , 41d , fluid ||| 1 0.0345864 1 0.000467132 2.718
、 ４１ ｄ 、 流体 貯留 ||| , 41d , fluid storing ||| 1 0.00048108 1 8.14764e-05 2.718
、 ４１ ｄ 、 流体 貯留 タンク ||| , 41d , fluid storing tank ||| 1 0.000284362 1 7.14906e-05 2.718
、 ６ および 流体 ||| and 6 and a fluid ||| 1 0.00410477 1 0.00964129 2.718
、 ６ および 流体 測 温 ||| and 6 and a fluid temperature measuring ||| 1 7.86306e-06 1 0.000721089 2.7
、 ６１ ’は 伝 熱 流体 ||| , 61 ’ a heat-transfer fluid ||| 1 0.00216443 1 0.000518098 2.718
、 ６１ は 伝 熱 流体 ||| , 61 is a heat-transfer fluid ||| 1 0.0185343 1 0.00119982 2.718
、 ７７ は 加工 液 ||| , 77 is a working fluid ||| 1 0.00864883 1 0.00028376 2.718
、 ７７ は 加工 液 ノズル ||| , 77 is a working fluid nozzle ||| 1 0.00764329 1 0.000210182 2.718
、 ７８ は 加工 液 ||| , and 78 is a working fluid ||| 1 0.00521192 1 1.09199e-05 2.718
、 Ｐ は 流体 ||| , p denotes fluid ||| 1 0.122825 1 0.000660778 2.718
、 Ｑ は 流体 の 流量 ||| , q denotes flow rate of fluid ||| 1 0.0145985 1 2.77106e-05 2.718
、 および 、 第 ２ 液 ||| and a second fluid ||| 0.333333 9.53798e-05 1 0.00106579 2.718
、 および 、 第 ２ 液 圧 ||| and a second fluid pressure ||| 0.5 3.19513e-05 1 0.00045599 2.718
、 この 液 ||| , and this fluid ||| 1 0.0557024 0.166667 0.000360561 2.718
、 この 液 ||| , the fluid ||| 0.0285714 0.0047175 0.166667 0.014136 2.718
、 この 液 ||| , this fluid ||| 1 0.0934985 0.333333 0.00959799 2.718
、 この 液 ||| 8 , and this fluid ||| 1 0.0557024 0.166667 8.11983e-08 2.718
、 この 液 圧 ||| , and this fluid pressure ||| 1 0.0186597 0.2 0.000154263 2.718
、 この 液 圧 ||| , the fluid pressure ||| 0.25 0.00158031 0.2 0.00604794 2.718
、 この 液 圧 ||| , this fluid pressure ||| 1 0.031321 0.4 0.00410641 2.718
、 この 液 圧 ||| 8 , and this fluid pressure ||| 1 0.0186597 0.2 3.47399e-08 2.718
、 この 液 圧 は ||| , and this fluid pressure is ||| 1 0.00536829 0.333333 4.16064e-05 2.718
、 この 液 圧 は ||| , the fluid pressure is ||| 0.5 0.000454647 0.333333 0.0016312 2.718
、 この 液 圧 は ||| 8 , and this fluid pressure is ||| 1 0.00536829 0.333333 9.36976e-09 2.718
、 この 液 圧 を ||| , this fluid pressure is ||| 1 0.00774038 1 0.000828222 2.718
、 この 作動 液 ||| , the hydraulic fluid ||| 0.125 0.000293068 1 0.000462624 2.718
、 この 磁性 流体 ||| , the magnetic fluid ||| 0.333333 0.00118769 0.5 0.0798598 2.718
、 この 磁性 流体 ||| the magnetic fluid ||| 0.0769231 0.000586521 0.5 0.239424 2.718
、 この 磁性 流体 ７３ ||| , the magnetic fluid 73 ||| 1 0.00108018 1 0.0687977 2.718
、 この 流体 ||| , each fluid ||| 1 0.00229816 1 0.000525922 2.718

句対応の例：自動車 を含む句の例

、 「 自動車 ||| , ‘‘ automobile ||| 1 0.142122 1 0.0659882 2.718
、 「 自動車 電話 ||| , ‘‘ automobile telephone ||| 1 0.0746708 1 0.0524542 2.718
、 「 自動車 電話 」 ||| , ‘‘ automobile telephone ’’ ||| 1 0.0258531 1 0.0386858 2.718
、 １０２ は 自動車 用 ||| , 102 is a vehicle ||| 1 0.000161113 1 0.00350135 2.718
、 １０２ は 自動車 用 エンジン ||| , 102 is a vehicle engine ||| 1 9.82163e-05 1 0.0031887 2.718
、 ３ は 自動車 ||| , 3 a mobile ||| 1 0.000371807 1 0.000168593 2.718
、 ３ は 自動車 電話 ||| , 3 a mobile telephone ||| 1 0.000195348 1 0.000134015 2.718
、 ３ は 自動車 電話 局 ||| , 3 a mobile telephone office ||| 1 7.2785e-05 1 3.45847e-06 2.718
、 ３ は 自動車 電話 局 、 ||| , 3 a mobile telephone office , ||| 1 4.96815e-05 1 1.31658e-06 2.718
、 ７ は 自動車 ||| , 7 a mobile ||| 1 0.000370597 1 0.000217899 2.718
、 ７ は 自動車 電話機 ||| , 7 a mobile telephone equipment ||| 1 5.71506e-05 1 4.29458e-07 2.718
、 ７ は 自動車 電話機 、 ||| , 7 a mobile telephone equipment , ||| 1 3.90098e-05 1 1.63488e-07 2.718
、 ＥＣ セル ６ を通じて 自動車 ||| vehicle through the ec cell 6 ||| 1 0.00011244 1 0.0167565 2.718
、 ＯＡ 機器 、 自動車 ||| , business machines , automobiles ||| 1 3.50735e-05 1 2.53658e-07 2.718
、 ＯＡ 機器 、 自動車 、 ||| , business machines , automobiles and ||| 1 4.58492e-06 1 9.529e-09 2.718
、 ＯＡ 機器 、 自動車 、 精密 ||| , business machines , automobiles and precision ||| 1 3.50085e-07 1 2
、 この よう な 電気 自動車 ||| , such an electric vehicle ||| 0.5 1.71207e-05 1 3.34838e-05 2.718
、 この よう な 電気 自動車 の ||| , such an electric vehicle ||| 0.5 4.51596e-06 1 3.34838e-05 2.718
、 この ハイブリッド 自動車 ||| , in the hybrid vehicle ||| 1 0.000761246 0.5 0.00269411 2.718
、 この ハイブリッド 自動車 ||| , in this hybrid vehicle ||| 1 0.0064154 0.5 0.00182924 2.718
、 この ハイブリッド 自動車 ５００ ||| , in the hybrid vehicle 500 ||| 1 0.00071914 1 0.0022104 2.718
、 この ハイブリッド 自動車 ５００ において ||| , in the hybrid vehicle 500 , ||| 1 2.0926e-05 1 0.000841
、 この ハイブリッド 自動車 ６００ ||| , in this hybrid vehicle 600 ||| 1 0.00591218 1 0.00136645 2.718
、 この ハイブリッド 自動車 ６００ において ||| , in this hybrid vehicle 600 , ||| 1 0.000172036 1 0.0005
、 しかも 自動車 ||| the automotive ||| 0.1 0.000244648 1 0.0106571 2.718
、 たとえば 自動車 用 ||| , for example , an automobile ||| 0.4 0.000664619 1 0.000540882 2.718
、 たとえば 自動車 用 エンジン ||| , for example , an automobile engine ||| 1 0.000405159 1 0.000492584
、 コードレス 電話 、 自動車 ||| , cordless telephones , car ||| 1 0.0338496 1 0.000443696 2.718
、 コードレス 電話 、 自動車 電話 ||| , cordless telephones , car telephones ||| 1 0.0157655 1 1.50196e-
、 ディーゼルエンジン 自動車 等 ||| diesel engine automobiles and the like ||| 1 0.0151152 1 3.01743e-05
、 ディーゼルエンジン 自動車 等 に ||| on diesel engine automobiles and the like ||| 1 0.00642812 1 2.24
、 ディジタル 自動車 電話 ||| a digital cellular ||| 0.2 2.39509e-05 1 0.000163602 2.718
、 ディジタル 自動車 電話 システム ||| a digital cellular system ||| 1 9.71561e-06 1 0.000148973 2.718
、 デイジタル 自動車 ||| the digital cellular ||| 1 5.80871e-05 1 0.00122759 2.718
、 デイジタル 自動車 電話 ||| the digital cellular telephone ||| 1 3.05191e-05 1 0.000975814 2.718
、 デイジタル 自動車 電話 ５０ ||| the digital cellular telephone 50 ||| 1 2.83228e-05 1 0.000844553 2.7
、 デイジタル 自動車 電話 ５０ において ||| the digital cellular telephone 50 , ||| 1 8.24154e-07 1 0.000
、 ハイブリッド 自動車 ||| a hybrid powered automobile ||| 0.5 0.0354085 0.333333 8.32655e-05 2.718
、 ハイブリッド 自動車 ||| the hybrid car ||| 0.166667 0.0441317 0.333333 0.0427639 2.718
、 ハイブリッド 自動車 ||| the hybrid vehicle ||| 0.0263158 0.0102417 0.333333 0.0694053 2.718
、 ハイブリッド 自動車 ５００ ||| the hybrid vehicle 500 ||| 1 0.00967523 1 0.056944 2.718
、 ハイブリッド 自動車 ５００ の ||| of the hybrid vehicle 500 ||| 1 0.00706622 1 0.0220125 2.718
、 ハイブリッド 自動車 ５００ の ヨー 方向 ||| yawing directions of the hybrid vehicle 500 ||| 1 0.00101
、 ハイブリッド 自動車 の ||| the hybrid car ||| 0.166667 0.0116407 1 0.0427639 2.718
、 ファクシミリ 、 自動車 ||| , facsimiles , automotive vehicles ||| 1 0.0560437 1 1.15431e-06 2.718
、 プラスチック 廃棄 物 、 自動車 ||| , plastic wastes , and automobile ||| 1 0.00690582 1 1.86092e-05 2

現行の句対応抽出の問題点

- 発見的な方法に基づいているため，改良がむずかしい

  発見的の意味：抜き出された句対応が良い句対応かどうかの評価が，機械翻訳実験による性能の変化でしか測定できない．確率的なモデルがない．確率的なモデルがあれば，最尤となる句対応を選べば，そのモデルの観点からは最良の句対応が得られる．

- しかし,「こういう句対応が良いのでは」といういくつかのモデルは，この発見的な方法と同等か少し性能が落ちる．

句の良さの評価

句単位に翻訳していく方法では，なるべく良い句対応を使った翻訳文を得たいので，そのためのスコアを定義する．一つの句対応について次の5つのスコアが良く使われる．

- 句翻訳確率 $\phi(\overline{f}|\overline{e})$
- 句翻訳確率 $\phi(\overline{e}|\overline{f})$
- 単語確率より $\mathrm{lex}(\overline{f}|\overline{e})$
- 単語確率より $\mathrm{lex}(\overline{e}|\overline{f})$
- 句ペナルティ $\exp(1)$ (固定値)

句翻訳確率 $\phi(\overline{f}|\overline{e})$

$$\phi(\overline{f}|\overline{e}) = \frac{\mathrm{count}(\overline{f}, \overline{e})}{\Sigma_{\overline{f}'}\,\mathrm{count}(\overline{f}', \overline{e})}$$

$$\mathrm{count}(\overline{f}, \overline{e}) = \text{句}\overline{f}\text{と}\overline{e}\text{が対応した回数}$$

$$\sum_{\overline{f}'}\mathrm{count}(\overline{f}', \overline{e}) = \text{句}\overline{e}\text{が出現した回数}$$

$\phi(\overline{f}|\overline{e})$は，句$\overline{e}$が出現したとき，それが$\overline{f}$に翻訳される確率と考えられる．$\phi(\overline{e}|\overline{f})$も同様である．

単語確率より $\mathbf{lex}(\overline{f}|\overline{e})$

$$\mathrm{lex}(\overline{f}|\overline{e}) = \max_{\mathbf{a}} P_w(\overline{f}|\overline{e}, \mathbf{a})$$

$$P_w(\overline{f}|\overline{e}, \mathbf{a}) = \prod_{i=1}^{n} E_w(f_i|\overline{e}, \mathbf{a})$$

$$E_w(f_i|\overline{e}, \mathbf{a}) = \frac{1}{\{j|(i,j) \in \mathbf{a}\}} \sum_{(i,j)\in\mathbf{a}} w(f_i|e_j)$$

$$w(f_i|e_j) = \text{単語対応の確率} = \frac{\mathrm{count}(f_i, e_j)}{\Sigma_{f'}\,\mathrm{count}(f', e_j)}$$

$E_w(f_i|\overline{e}, \mathbf{a}) = w(f_i|e_j)$の$e_j$に関する平均値

$P_w(\overline{f}|\overline{e}, \mathbf{a}) = E_w(f_i|\overline{e}, \mathbf{a})$の積. f_iの条件付き独立を仮定

$\mathrm{lex}(\overline{f}|\overline{e}) = P_w(\overline{f}|\overline{e}, \mathbf{a})$が最大の$\mathbf{a}$に関する確率を採用

句ペナルティ

$\exp(1)$に固定．スコアとしては，対数を利用するので，$\log(\exp(1)) = 1$となる．
翻訳に使われた句の$\log$(句ペナルティ)の和は，翻訳に使われた句の数である．
実際の翻訳のスコアには，この句ペナルティ(翻訳に使われた句の数)に，ある重みλを掛けたものを利用する．
このλは自動推定する．
もし，$\lambda > 0$なら，句の数が多いほど翻訳候補のスコアは良くなる．
$\lambda < 0$なら，句の数が少ないほど良い．
通常は $\lambda < 0$で，少ない方が良い．これは同じ単語数からなる文であれば，使われた句の数が少ない (句の長さは長い)方が良いことになる．

フレーズ抽出まとめ

発見的手続きを使うことにより，日英のフレーズ対応を得て，そこに各種のスコアを付けることができることをみた.

問題 (15分)

自動作成したフレーズテーブルには，さまざまな誤りが含まれる．そのようなフレーズテーブルの中から正しそうなフレーズのみを自動抽出するには，どうしたら良いか．

回答例1

	$\overline{f}$	$\overline{f}$ 以外
$\overline{e}$	a	b
$\overline{e}$以外	c	d

$$
\begin{aligned}
a &= \overline{e}と\overline{f}の対応数 \\
b &= \overline{e}の出現数 - a \\
c &= \overline{f}の出現数 - a \\
d &= 句の対応の総数 - a - b - c
\end{aligned}
$$

とすると $\frac{a}{c} \gg \frac{b}{d}$ のとき，$\overline{e}$ と $\overline{f}$ は良い対応と考えられる．これは，$\overline{e}$ と $\overline{f}$ の独立性の検定をして，独立性の低いものが良いということなので，その検定結果により，独立性の低い順に句対応をソートして，上位のみを利用する．

回答例2

- パラレルコーパスを2つに分ける.
- それぞれのコーパスから1つずつフレーズテーブルを作る
- 両方のフレーズテーブルに含まれるようなフレーズのみを採用する

これらの回答例の得失

得 ある程度は信頼性の高いフレーズテーブルができる

失 カバレッジが低くなる

→ 得が大きいか，失が大きいかは実験してみないと分からない.
今のところ，翻訳の性能を下げずに，ノイズ除去ができるらしいということが分かっている.