12. 対訳コーパスの自動作成

内山将夫@NICT
mutiyama@nict.go.jp

対訳コーパスの自動作成

対訳関係にある文章から
対訳関係にある文対応を同定することにより
対訳コーパスを自動作成する

対訳文対応同定の手順

1. 英語文章 E と日本語文章 J を用意する
2. E を辞書引きして，日本語単語に変換する
3. J を単語に分割する
4. 単語同士の対応を利用して，文同士の対応をみつける

日本語テキストの例

j0 天文学者

j1 ある天文学者は、夜になるとしょっちゅう、星を観測しに郊外へと出かけて行った。

j2 ある晩、彼は、星に気を取られていて、誤って深い井戸に落ちてしまった。

j3 彼は、打ち身や切り傷をつくって、悲鳴を上げた。

j4 その声を聞きつけて、近所の人が井戸へと飛んできた。

j5 そして何が起きたのかを知ると、こんな事を言った。

j6 「天国を覗くことばかりに、うつつを抜かしてないで、少しは足下に注意を払いなさいな」

英語テキストの例

e0 The Astronomer

e1 AN ASTRONOMER used to go out at night to observe the stars.

e2 One evening, as he wandered through the suburbs with his whole attention fixed on the sky,

e3 he fell accidentally into a deep well.

e4 While he lamented and bewailed his sores and bruises,

e5 and cried loudly for help,

e6 a neighbor ran to the well,

e7 and learning what had happened said:

e8 "Hark ye, old fellow, why, in striving to pry into what is in heaven,

e9 do you not manage to see what is on earth?

e10 "

日本語テキストを単語に分割し内容語をとる

```
j0 天文学 天文学
j0 者 者
...
j1 出かけ 出かける
j1 行っ 行く
j2 晩 晩
j2 彼 彼
...
j2 井戸 井戸
j2 落ち 落ちる
j3 彼 彼
...
j3 上げ 上げる
j4 声 声
j4 聞きつけ 聞きつける
j4 近所 近所
...
j5 起き 起きる
j5 知る 知る
j5 事 事
...
j6 注意 注意
j6 払い 払う
```

英語テキストを辞書引きする

e0 astronomer 者 天文学 astronomer
e1 go_out_at_night 戸出 夜 夜歩き
e1 observe マーク 意見 監視 観ずる 観る ...
e1 stars 星屑 星辰 天涯孤独 到達 不可能 ...
....
e2 fixed あてがう こわばる 一定 確固たる...
e2 sky お天気 たる ひばり スカイ ...
e3 fell すさまじい たくましい ぴりっと ...
...
e4 lamented 哀悼 故人 死者 惜しむ ...
e4 bewailed bewail 哀 号 号泣 愁傷 ...
e4 sores sore しゃくにさわる ただれる ...
...
e6 neighbor 近く 近所 近付ける 近傍 ....
e6 ran くつ下 ぶつける バス 引く ...
e6 well い いい す たんまり ...
e7 learning 憶 憶える 科学 会釈 ...
...
e8 striving 合 辛苦 戦 張 張りあい ...
e8 pry こじあける せんさく てこ ...
...
e9 see お目もじ ご覧 しばしば ...
...

文同士の対応をみつける

天文学者
The Astronomer
ある天文学者は、夜になるとしょっちゅう、星を観測しに郊外へと出かけて行った。
AN ASTRONOMER used to go out at night to observe the stars.
ある晩、彼は、星に気を取られていて、誤って深い井戸に落ちてしまった。
One evening, as he wandered through the suburbs with his whole attention fixed on the sky,
he fell accidentally into a deep well.
彼は、打ち身や切り傷をつくって、悲鳴を上げた。
While he lamented and bewailed his sores and bruises,
その声を聞きつけて、近所の人が井戸へと飛んできた。
and cried loudly for help,
a neighbor ran to the well,
そして何が起きたのかを知ると、こんな事を言った。
and learning what had happened said:
「天国を覗くことばかりに、うつつを抜かしてないで、少しは足下に注意を払いなさいな」
”Hark ye, old fellow, why, in striving to pry into what is in heaven,
do you not manage to see what is on earth?
”

どういう文同士の対応をみつけるか

可能な対応例はたくさんある
そのうちで最適なものを見つけたい

j0	e0
j1	e1
j2	e2, e3
j3	e4
j4	e5,e6
j5	e7
j6	e8,e9,e10

j0	e0,e1
j1,j2	e2, e3
j3	e4
j4	e5,e6,e7
j5,j6	e8,e9,e10

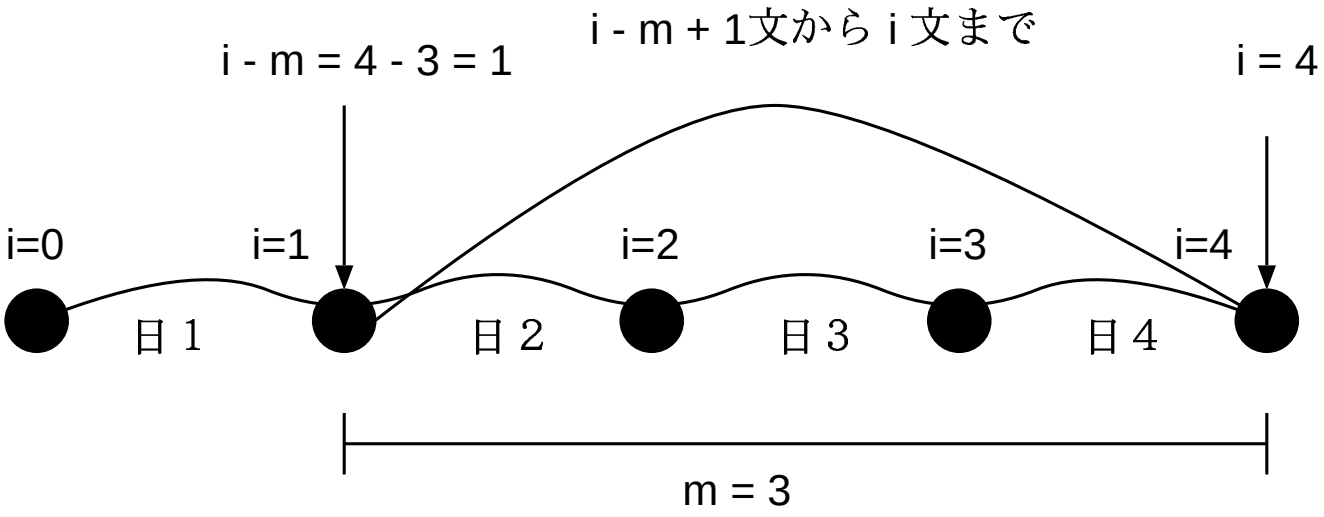
最適性の定義

$$\arg \max_{\text{文対応例} \in \text{可能な文対応集合}} \sum_{\text{各文対応}} \text{SIM(各文対応)}$$

- いくつもの文対応例があるが，そのうち，上式を満すものをとる
- このとき，SIMは，ある文対応例における，各文対応の類似度である.
- よって，上式における和は，ある文対応例のスコアとして,
- 各文対応の類似度の和を採用している.

要するに，類似度最大となるような文対応例を求める

数式表現



まず，複数文を表現するために，iが日文$_i$と日文$_{i+1}$の間にあるとすると

$$\text{日}(i-m,i) = \text{日}_{i-m+1}\text{日}_{i-m+2}\ldots\text{日}_{i}$$

$$\text{英}(j-n,j) = \text{英}_{j-n+1}\text{英}_{j-n+2}\ldots\text{英}_{j}$$

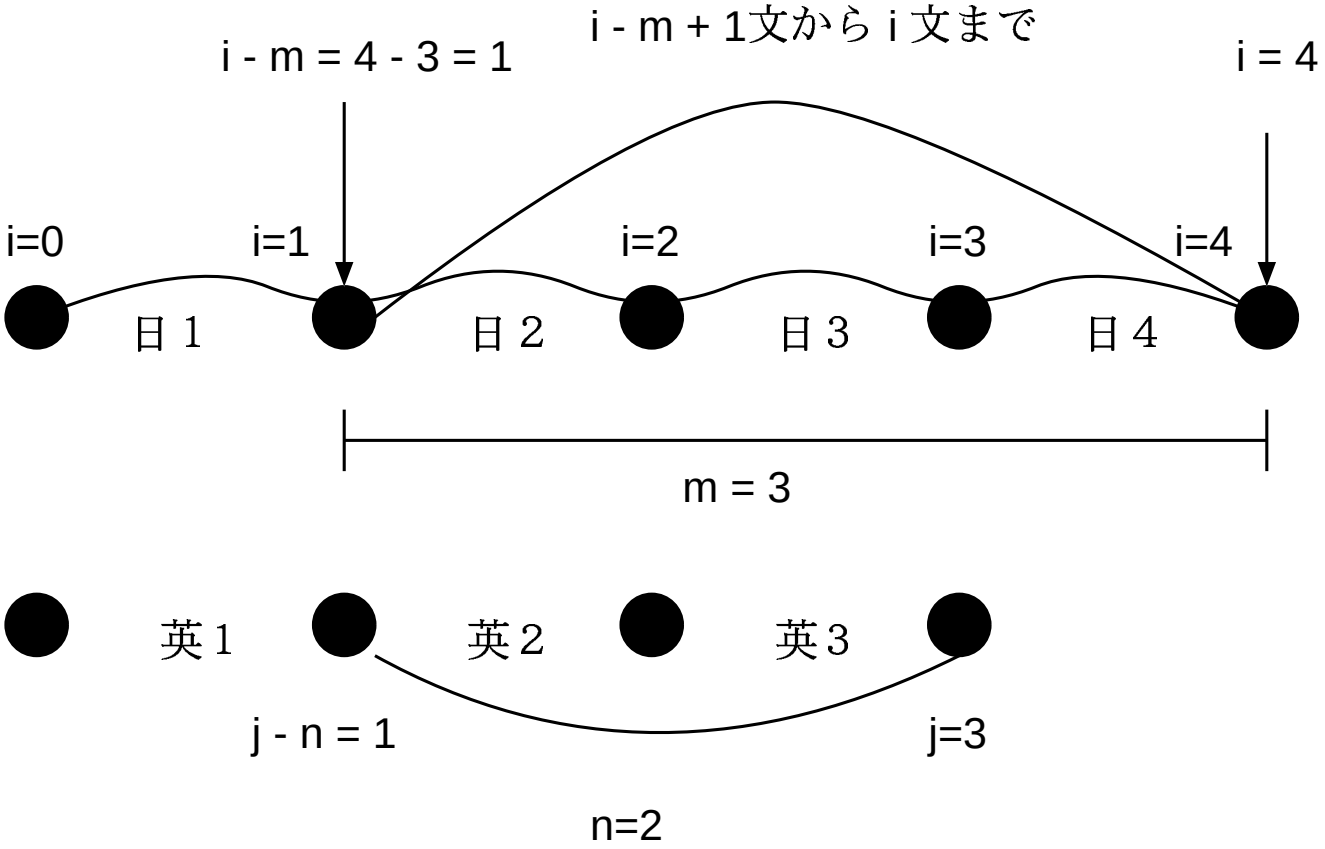
は，日$_i$以前のm単語と，英$_j$以前のn単語である．そうするとJを日本語文数，Eを日本語文数として，

$$\max \sum_{i=0}^{J} \sum_{j=0}^{E} \sum_{m=0}^{i} \sum_{n=0}^{j} \mathrm{SIM}(\text{日}(i-m,i), \text{英}(j-n,j))$$

となるような類似度の中で，ちゃんと対応の整合性がとれているようなものをとる．

再帰式による最大スコアの計算

今計算したいものは，$S(i,j)$である．
$S(i,j)$とは，日本語文がi文，英語文がj文並べられたときの最適スコアである．



このとき，$S(i-m,j-n)$が既に求まっていると仮定する．つまり，日本語の$i-m$文までと，英語の$j-n$文までは，対応済みとする．上例だと,「日$_1$日$_2$日$_3$日$_4$」と「英$_1$英$_2$英$_3$」のスコアを求めるとき，$m=3,n=2$のときには，日$_1$と英$_1$のスコアは計算済とする．すると

$$S(i,j) = S(i-m,j-n) + \mathrm{SIM}(\text{日}(i-m,i) + \text{英}(j-n,j))$$

により$S(i,j)$が求まる．

つまり

$$\begin{aligned} & S(\text{日}_1\text{日}_2\text{日}_3\text{日}_4,\ \text{英}_1\text{英}_2\text{英}_3) \\ & = S(\text{日}_1,\ \text{英}_1) + \mathrm{SIM}(\text{日}_2\text{日}_3\text{日}_4, \text{英}_2\text{英}_3) \qquad (1) \end{aligned}$$

である．これは，m と n が与えられている場合であるが，m と n については，固定されていないので，全ての m と n を考える必要がある．ただし，あまり多くすると計算量が多くなるので，たとえば，

$$(n, m) \in \{(1,0), (1,1), (1,2), (1,3), (0,1), (2,1), (3,1), (2,2)\}$$

とする．そのとき，

$$S(i,j) = \max_{m,n} S(i-m, j-n) + \mathrm{SIM}(\text{日}(i-m, i), \text{英}(j-n, j))$$

である．また，

$$S(0,0) = 0$$

なので，帰納的に最適スコアを定義できる．

$S(i,j)$の計算法

テーブルSを用意して，それを小さい方から埋めていく

$$日_1(a,b), 日_2(c), 日_3(d)$$

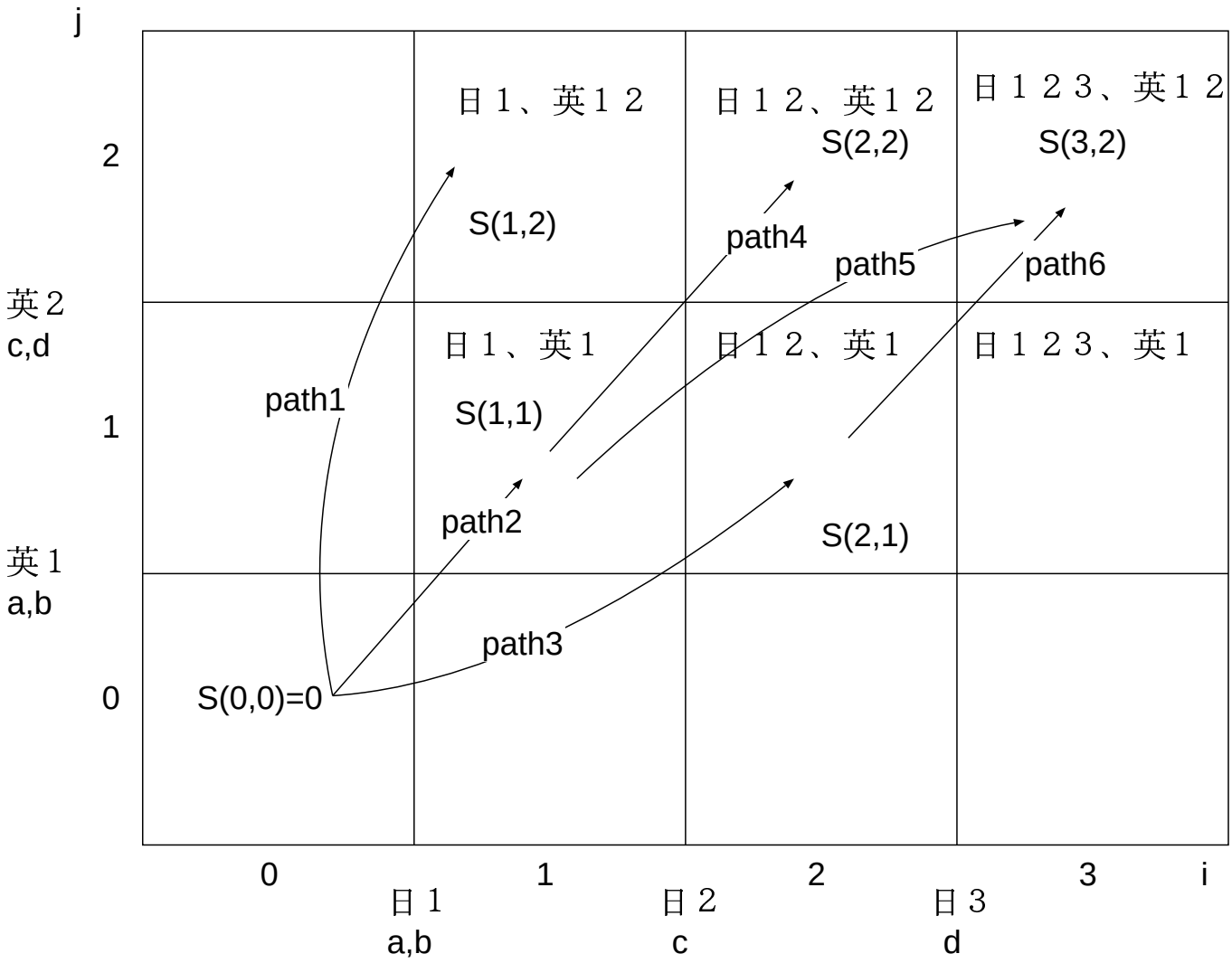
$$英_1(a,b), 英_2(c,d)$$

について，1対1，1対2，2対1の対応を許すときに，どのような文対応があるかをみる．a,b,c,dは単語とする．類似度は

$$\mathrm{SIM}(J,E) = \frac{2|J \cap E|}{|J| + |E|}$$

により計算する．ただし，J,Eは日英における単語集合．

テーブル S


j
2
1
0
英２
c,d
英１
a,b
日１、英１２
日１２、英１２
日１２３、英１２
S(2,2)
S(3,2)
S(1,2)
path4
path5
path6
日１、英１
日１２、英１
日１２３、英１
path1
S(1,1)
path2
S(2,1)
path3
S(0,0)=0
0
1
2
3
i
日１
a,b
日２
c
日３
d

テーブル S の埋め方

- まず，$S(0,0)=0$ とする

次に，$i,\ j$ の小さい方から $S(i,j)$ を埋めていく．このとき，1対1，1対2，2対1の文対応のみを許すので，図のパスのように，斜め方向に進む対応のみが許される．
したがって，
$i=1$ のときには，

path1 $\mathrm{SIM}(\text{日}\,1,\text{英}\,12) = \mathrm{SIM}(\{a,b\},\{a,b,c,d\}) = \frac{2\times2}{2+4} = \frac{4}{6} = 0.67$

path2 $\mathrm{SIM}(\text{日}\,1,\text{英}\,1) = \mathrm{SIM}(\{a,b\},\{a,b\}) = \frac{2\times2}{2+2} = 1$

より

$$
\begin{aligned}
S(1,1) &= S(0,0) + \mathrm{SIM}(\text{日}\,1,\text{英}\,1) = 1 \\
S(1,2) &= S(0,0) + \mathrm{SIM}(\text{日}\,1,\text{英}\,12) = 0.67
\end{aligned}
$$

$i = 2$のときには，

path3 $\mathrm{SIM}(\text{日}12, \text{英}1) = \mathrm{SIM}(\{a, b, c\}, \{a, b\}) = \frac{2\times2}{3+2} = \frac{4}{5} = 0.8$

path4 $\mathrm{SIM}(\text{日}2, \text{英}2) = \mathrm{SIM}(\{c\}, \{c, d\}) = \frac{2\times1}{1+2} = \frac{2}{3} = 0.67$

path1 から S(2,2) に移るためには，真横に移動しないといけないが，これは，日本語文は消費されるが，英語文は消費されないので，1対0 の対応となる．この対応は許されていない．よって，

$$S(2,1) = S(0,0) + \mathrm{SIM}(\text{日}12, \text{英}1) = 0.8$$
$$S(2,2) = S(1,1) + \mathrm{SIM}(\text{日}2, \text{英}2) = 1 + 0.67 = 1.67$$

$i = 3$のときには，

path5 $\mathrm{SIM}(\text{日}23, \text{英}2) = \mathrm{SIM}(\{c, d\}, \{c, d\}) = \frac{2\times2}{2+2} = 1$

path6 $\mathrm{SIM}(\text{日}3, \text{英}2) = \mathrm{SIM}(\{d\}, \{c, d\}) = \frac{2\times1}{1+2} = \frac{2}{3} = 0.67$

よって，path5 を通ったときには

$$S(3, 2) = S(1, 1) + \mathrm{SIM}(\text{日}23, \text{英}2) = 1 + 1 = 2$$

path6 を通ったときには

$$S(3, 2) = S(2, 1) + \mathrm{SIM}(\text{日}3, \text{英}2) = 0.8 + 0.67 = 1.47$$

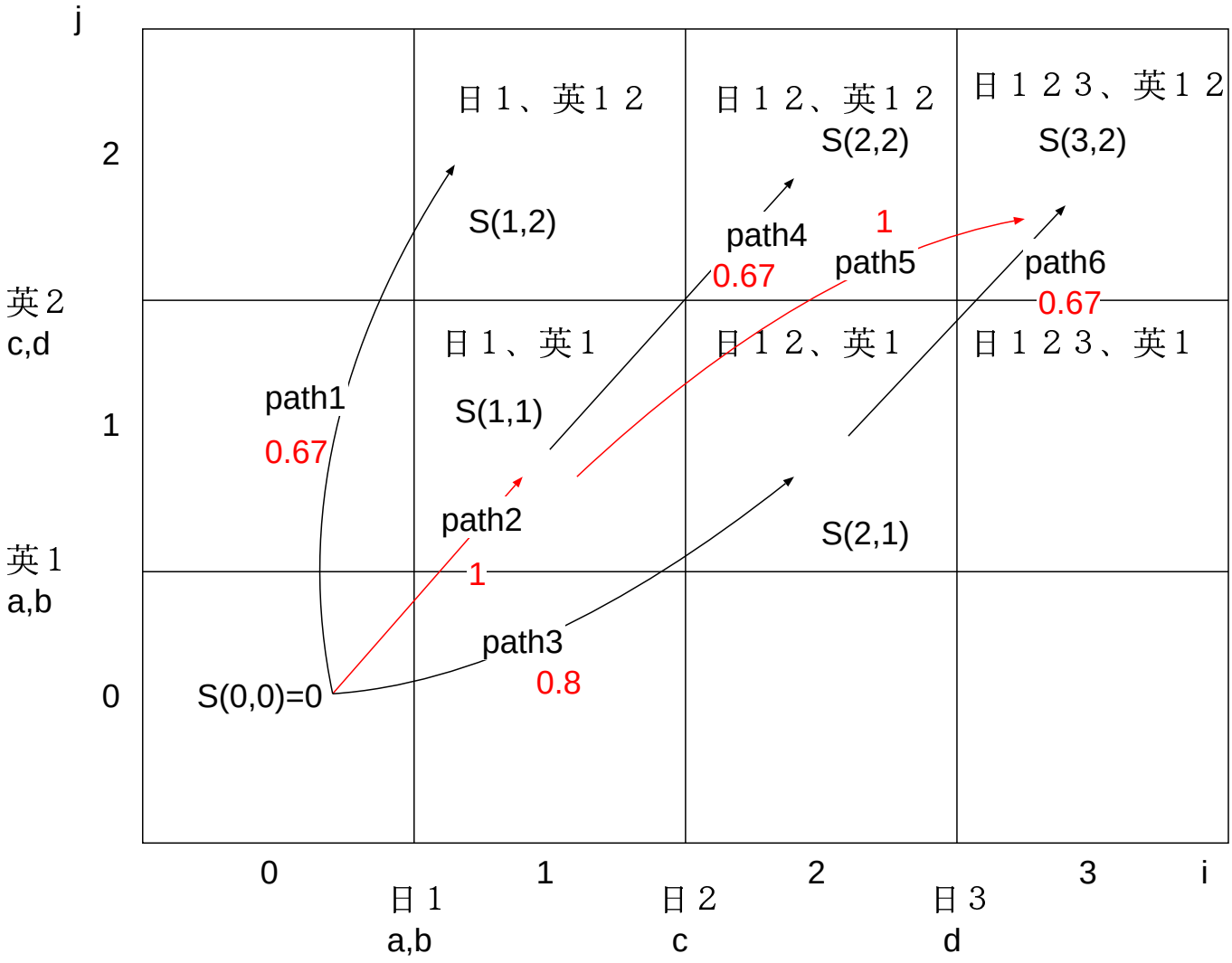
これらの最大値をとって，

$$S(3, 2) = 2$$

$S(0,0)$から path2 と path5 を通って $S(3,2)$ に至るパスが最適解である．そのときの対応は，

$$\begin{aligned} S(3,2) &= S(1,1) + \mathrm{SIM}(\text{日}23, \text{英}2) \\ &= S(0,0) + \mathrm{SIM}(\text{日}1, \text{英}1) + \mathrm{SIM}(\text{日}23, \text{英}2) \\ &= 0 + 1 + 1 = 2 \end{aligned}$$

より，「日1と英1」「日23と英2」が対応している．



文対応のまとめ

日本語文数をJ，英語文数をEとすると

$$S(J, E)$$

は最大スコアによる文対応のスコアとなる
このとき，各文対応を$a_1, a_2, ..., a_N$とすると

$$S(J, E) = \sum_{i=1}^{N} \mathrm{SIM}(a_i)$$

ただし，a_iは，前述の path にあたると考えてよく，$\mathrm{SIM}(a_i)$は，対応a_iを構成する日本語文と英語文の類似度である.
このとき，

$$\mathrm{AVSIM} = \frac{S(J, E)}{N}$$

は，各対応a_iの類似度SIMの平均値である.

対訳コーパスの自動構築

新聞記事のように，必ずしも直訳されていない対訳テキスト$T_1, T_2, ...$があるとする．
各 T_i について，文対応を求めると，それにより$\mathrm{AVSIM}(T_i)$が求まる．
$\mathrm{AVSIM}(T_i)$が大きいT_iは良く似た文対応からなると考えられる．
したがって，この値の大きいテキストをとることにより，対訳の度合が高いテキストを取ることができる．
また，良い対応のテキストに含まれる文は，良い文対応だと考えられるので，

$$\text{文対応のスコア} = \mathrm{SIM} \times \mathrm{AVSIM}$$

は，テキストの類似度までを考慮した文対応スコアである．
この文対応スコアが大きいものから対訳コーパスに採用する．

対訳コーパスについてのまとめ

- 現状で利用可能な対訳コーパスとして，NICTで公開しているものを紹介した．
- 対訳コーパスは機械翻訳以外にも使えることを示した．
- 対訳テキストから対訳コーパスを自動作成する方法を示した．