Technical Reports

ガチャ align

Sentence Alignment

Task:

- To align sentences to its corresponding translation

Assumption:

- Sentence alignment can be done monotonically

Challenges:

- non-1:1 alignments, insertions, deletions, incomplete translations

Assumption: Sentence alignment can be done monotonically

On the third day of this much-anticipated exhibition , the event will be open to the general public as a ticketed shopping event called the Blueprint Emporium .	
Held in collaboration with Zouk , be prepared to be seen at this chic fashion party on 1 May at the F1 Pit Building , and take home limited edition and past season samples from Asian designers and labels that cannot be found anywhere else .	この 大き な 期待 が さ れ て いる 展示 会 の 3 日 目 に は 、 この イベント は 一般 市民 に対し て も チケット 制 の イベント として 開か れ 、 また 、 ブループリント 百貨 店 と 呼ば れる 買い 物 イベント が 、 伝説 的 な 地元 の ナイトクラブ ・ ズーク と の 共同 で 開か れ ます 、 最先 端 の ファッション の 才能 が 披露 さ れ ます 。

(*black text doesn't get translated)

Challenges: non-1:1 alignments, insertions, deletions, **incomplete translations**

Aiming to attract about 3000 visitors , the Blueprint Emporium allows exhibitors to test their brands on the Singapore market , as this event will showcase offerings from cutting-edge brands , many of which are not available commercially in Singapore .	一晩 だけ でも 、 予想 観客 動員 数 の ３ ０ ０ 人 は 多目 的 の 、 ５ ０ 人 の デザイナー と ラウル （ RAOUL ）、 オールドレスアップ （ alldressedup ） や ウィキッド · ソング （ Wykidd Song ） など の デザイナー や 一流 地元 ブランド を 含む 国内 外 の 市場 から の 集団 を 収容 する こと が できる ファッションクラブ 空間 で 独特 に 圧倒 的 な 魅力 の ある 体験 に 迎え られる でしょ う 。

Sentence Alignment Approaches

Lexical methods

- corresponding sentences contain more corresponding words

Length-based methods

- sentences that correspond to each other are also similar in length (characters or words)

Combined methods

- use lexical cues in length-based settings

Gale-Church's Length Based model

- Define a distance based on the costs of aligning source to target sentences (for a fixed finite set of possible alignment types)

- Minimize this distance by finding the best alignment using dynamic programming → recursive definition of

$$D(i,j) = min \begin{cases} D(i,j-1) & + & cost(align_{0:1}, 0, t_j) \\ D(i-1,j) & + & cost(align_{1:0}, s_i, 0) \\ D(i-1,j-1) & + & cost(align_{1:1}, s_i, t_j) \\ D(i-1,j-2) & + & cost(align_{1:2}, s_i, t_{j-1}..t_j) \\ D(i-2,j-1) & + & cost(align_{2:1}, s_{j-1}..s_j, t_j) \\ D(i-2,j-2) & + & cost(align_{2:2}, s_{j-1}..s_j, t_{j-1}..t_j) \end{cases}$$

Gale-Church Cost Function

- assume that each character in the source language generates ***c*** characters in the target language with variance, ***s2*** and distance function:
- distance **δ = (*srclen – trglen* c*)/ *sqrt*(*srclen*s2*)** is normally distributed and **P(δ|*aligntype*)** gives the probability of observing a specific length-pair
- define prior probabilities of **P(*aligntype*)**

→ finally the cost function:

_log_P(_aligntype_)P(δ|_aligntype_)

Gale-Church Algorithm

- compute alignment costs for each sentence pair (i, j)
- start with 0^{th} source, 0^{th} target sentences and fill the entire table
- read the alignment path with minimal costs

Gale-Church Parameters

- empirically find parameters ***c, s2*** and **P(*aligntype*)** from example corpora
- Gale-Church used a German-English corpus and defined ***c = 1*** , ***s2 = 6.8***

P(*aligntype = 1 : 1*) = *0.89*

P(*aligntype = 1 : 0*) = *0.0099*

P(*aligntype = 0 : 1*) = *0.0099*

P(*aligntype = 2 : 1*) = *0.089*

P(*aligntype = 1 : 2*) = *0.089*

P(*aligntype = 2 : 2*) = *0.011*

Gale-Church Tweak

- fixed parameters are based on German-English corpus but it works surprisingly well for most European language pairs.

- **What if we apply Gale-Church to non-European Languages?**
- **What if we tweak the fixed parameters?**

ガチャalign (GaCha align)

- An experiment to test how parameters, {***c*, *s2*** and ***P(aligntype)***} affects the accuracy of alignment for an English-Japanese corpus

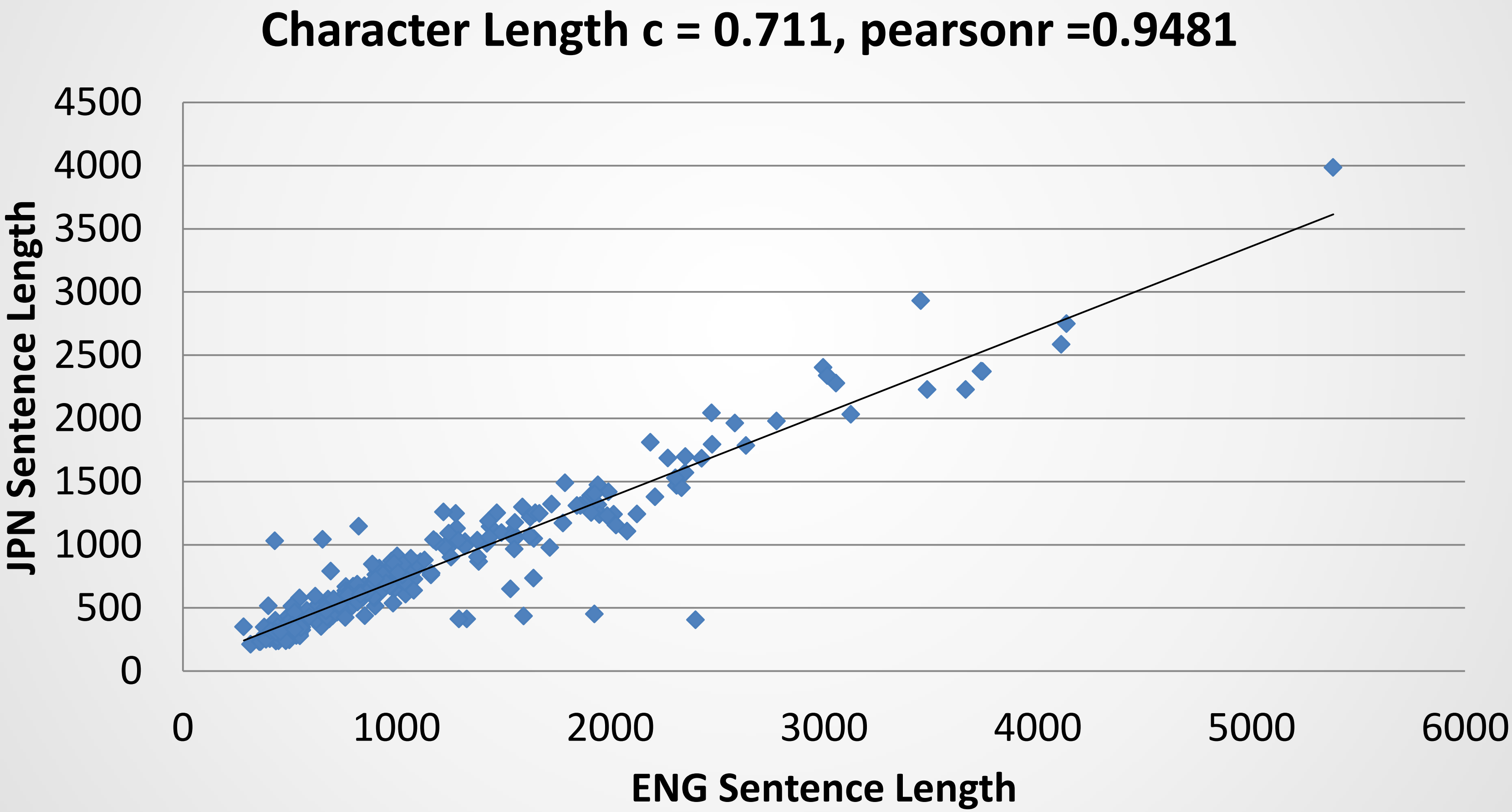
Tasks:

- Run Gale-Church alignments
  - with calculated parameters from the corpus
  - to determine what is the optimal value for the c and s2 for best accuracy

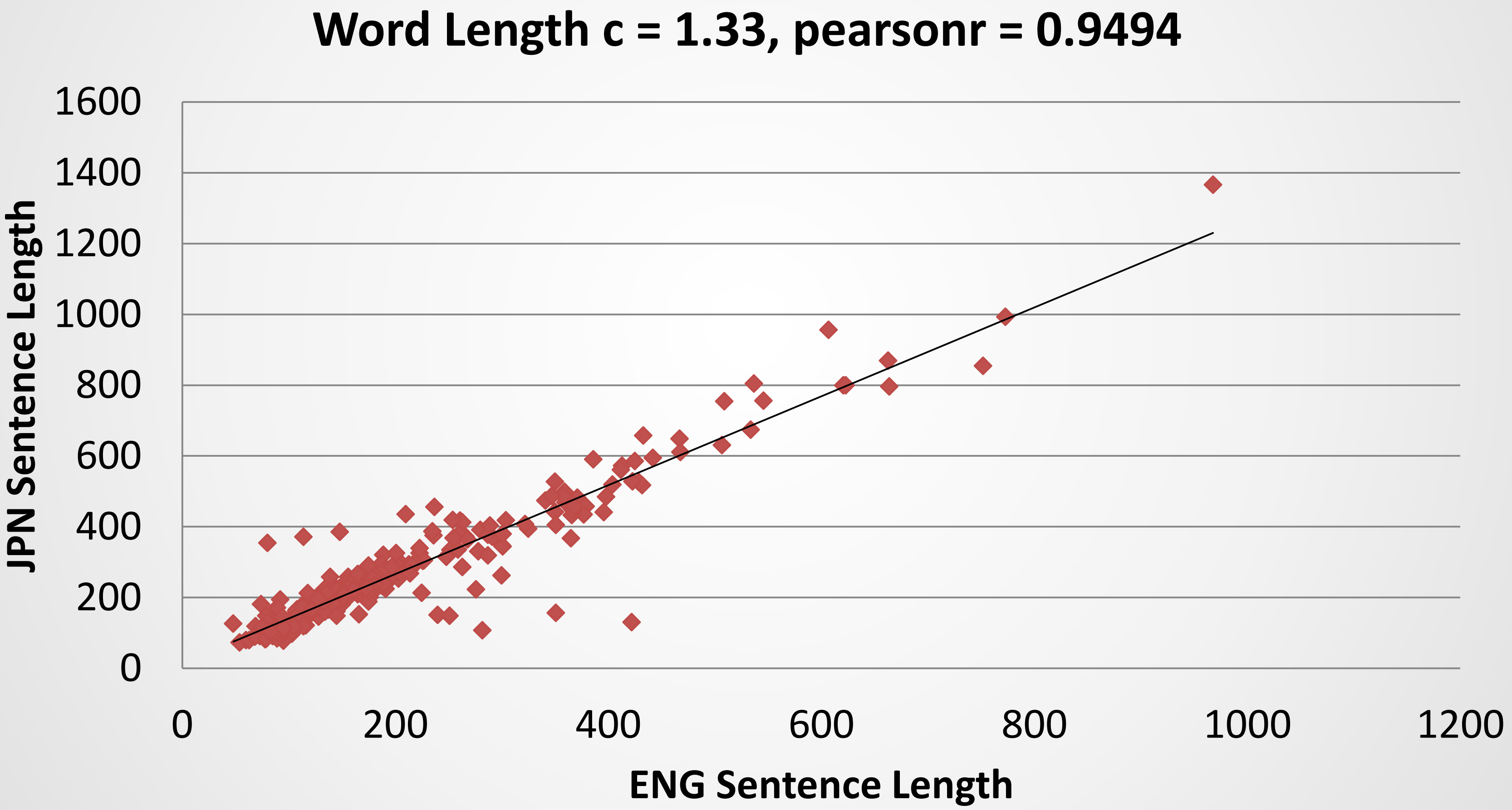
Data

- 1853 human-aligned ENG-JPN sentences from NTU-MC

ガチャalign


Character Length c = 0.711, pearsonr =0.9481
4500
4000
3500
3000
2500
2000
1500
1000
500
0
JPN Sentence Length
0
1000
2000
3000
4000
5000
6000
ENG Sentence Length

ガチャalign



Word Length c = 1.33, pearsonr = 0.9494
1600
1400
1200
1000
800
600
400
200
0
JPN Sentence Length
0
200
400
600
800
1000
1200
ENG Sentence Length

Results with calculated c and s2

	Char-based	Word-based
c (mean)	0.711	1.332
s2 (variance)	416.89	77.64

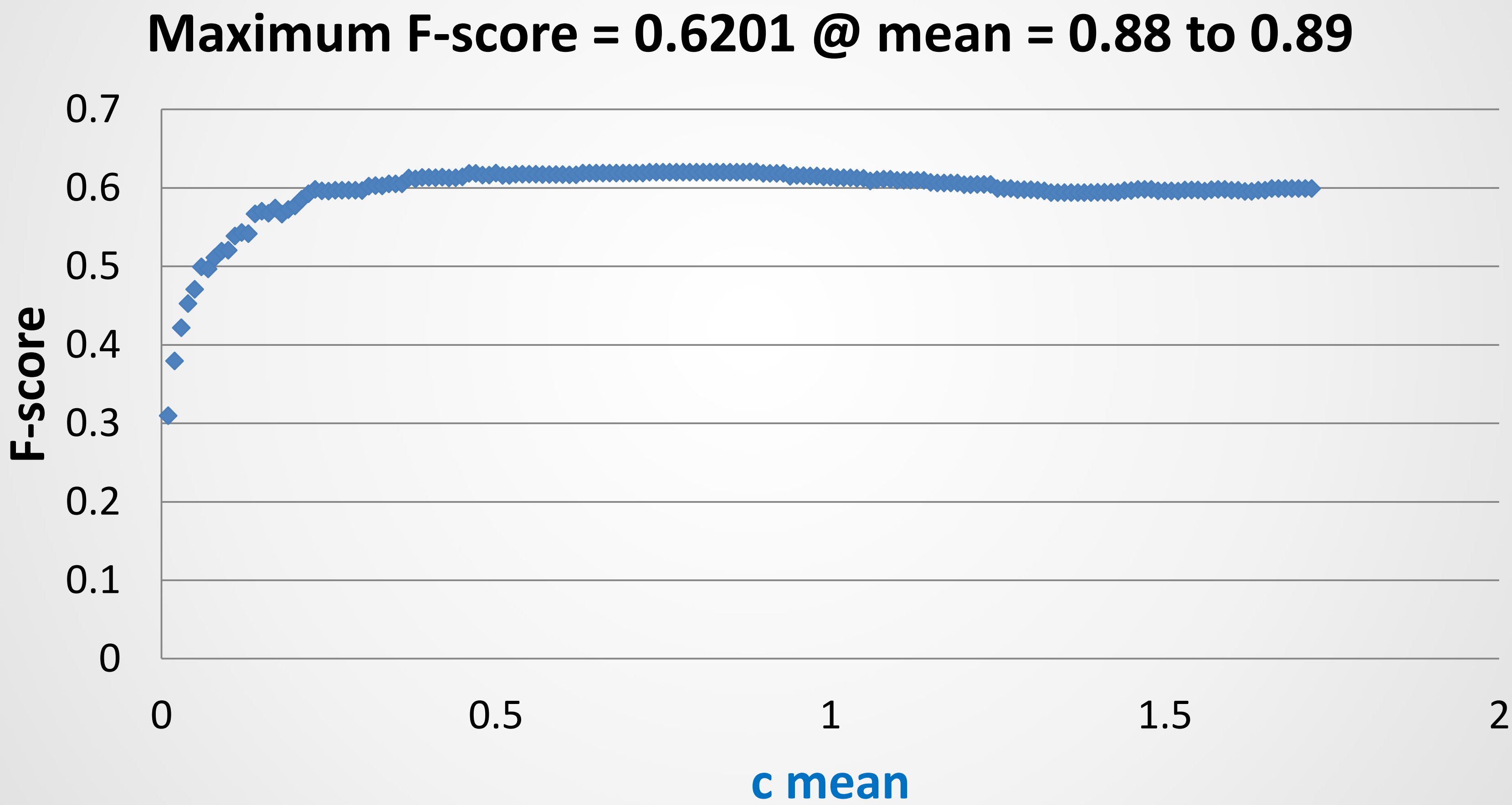
ガチャalign (c, s2 tweak)

char-based	c	S2	Recall	Precision	F-score
Default	1	6.8	0.6657	0.5637	0.6104
c tweak	**0.711**	6.8	**0.6724**	**0.5731**	**0.6188**
s2 tweak	1	416.89	0.6224	0.5206	0.5670
c + s2 tweak	0.711	416.89	0.5758	0.4868	0.5276
word-based	**c**	**s2**	**Recall**	**Precision**	**F-score**
c tweak	1.33	6.8	0.6413	0.5531	0.5940
s2 tweak	1	77.64	0.6552	0.5483	0.5970
c + s2 tweak	1.33	77.64	0.6618	0.5602	0.6068

ガチャalign

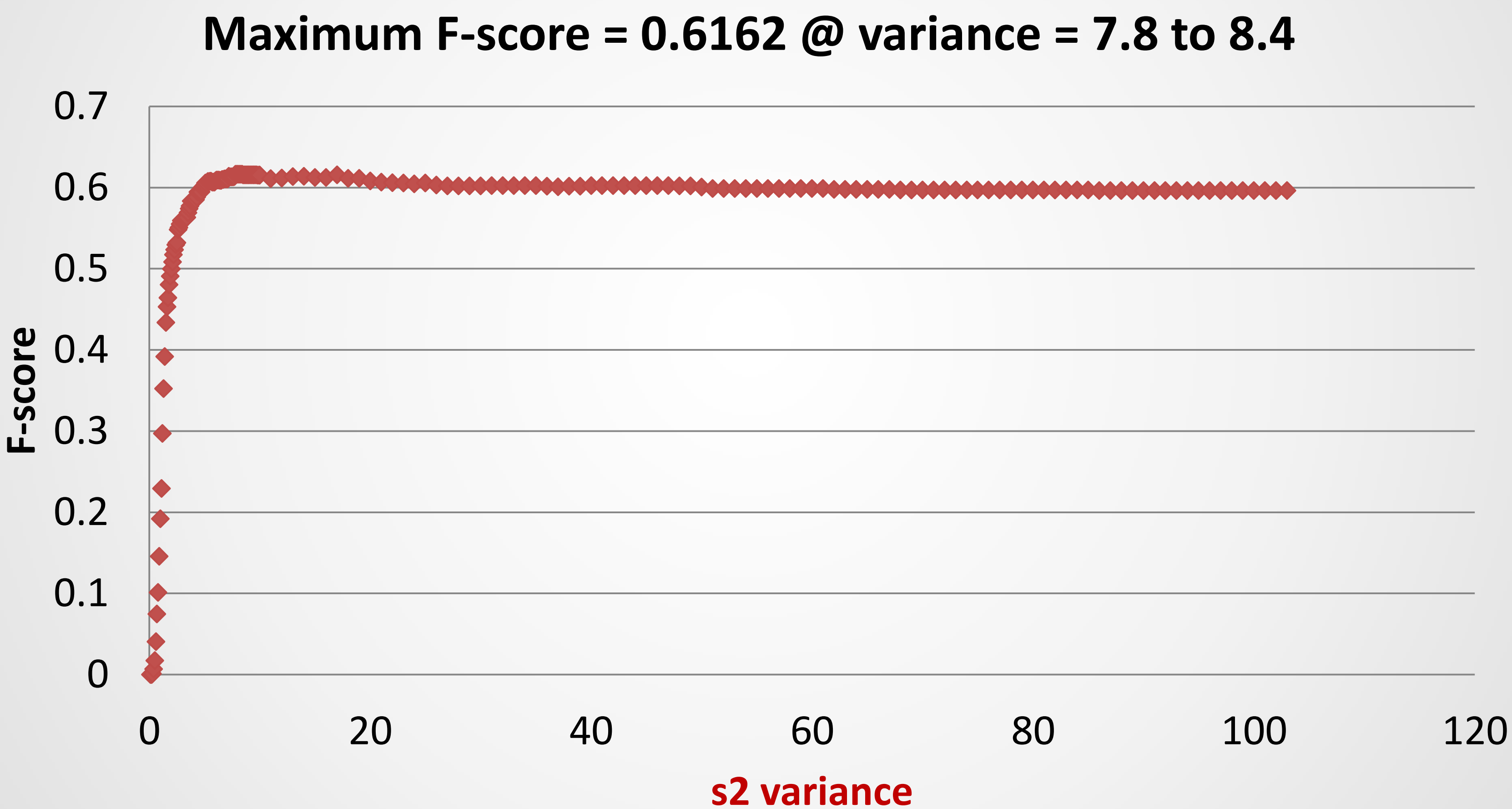
	Original Gale-Church	Calculated from NTU-MC (eng-jpn)
P(aligntype = 1 : 1)	0.89	0.72153
P(aligntype = 1 : 0)	0.0099	0.05288
P(aligntype = 0 : 1)	0.0099	0.03022
P(aligntype = 2 : 1)	0.89	0.00702
P(aligntype = 1 : 2)	0.89	0.16352
P(aligntype = 2 : 2)	0.11	0.00216
**P(aligntype = 1 : 3)*	-	0.01619
**P(aligntype = 1 : 4)*	-	0.01619
**P(aligntype = others)*	-	0.00377
F-score with c =0.711	**0.6188**	0.6243 , ***0.6091**

ガチャalign



Maximum F-score = 0.6201 @ mean = 0.88 to 0.89
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
F-score
0
0.5
1
1.5
2
c mean

ガチャalign



Maximum F-score = 0.6162 @ variance = 7.8 to 8.4
F-score
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120
s2 variance

ガチャalign

Conclusion

- Gale-Church algorithm seems to be robust enough to be unaffected by language specific ***length proportion*** or ***alignment types***

- Tweaking ***character mean*** between the src and trg text can be *done simply* and we have shown that it improves accuracy.
  - *default:* ***c=1.00****,* ***s2=6.8****, f-score =* ***0.6104***
  - *calculated mean:* ***c=0.71****, f-score =* ***0.6188***
  - *optimal:* ***c=0.88****,* ***s2=7.8****, f-score =* ***0.6199***

- Using text dependent ***alignment types probabilities*** don't affect performance too
  - default: f-score = ***0.6199*** tweaked: f-score = ***0.6290***

ガチャalign

Future Works

- Hybrid model by
  - using a dictionary and adding a weight to the length based:

$$dic\_weight * \log P(aligntype) P(\delta \mid aligntype)$$

English	Japanese
Char kway teow , loosely translated as " stir-fried rice cake strips " , is made by stir-frying flat rice noodles (similar to the Italian tagliatelle) with light and dark soy sauce , a dash of belachan (shrimp paste) , tamarind juice , bean sprouts , Chinese chives , lap cheong (Chinese sausages) and cockles .	チャー・クウェイ・ティオは、“炒めた平たい米麺”を意味します。 炒めた平たい米麺（イタリア料理のタリアテーレに似ている）を、薄味の醤油や少量のベラチャン（エビのすり身）、タマリンドジュース、もやし、ニラ、ラプチョン（腸詰）、ザルガイなどと伴に炒めて作ります。
In its original recipe , the rice noodles are also stir-fried in pork fat using crisp bits of pork lard , resulting in a distinctively rich taste .	オリジナルのレシピでは、少量のポークラードや豚脂を使って米麺も一緒に炒め、格別に濃厚な味にします。
In recent years , the dish as evolved into a healthier version with hawkers serving up more greens and adding less oil .	昨今の屋台では、緑黄色野菜をより多く用いて油分も控えめにした、より健康的なものを提供しています。
Char kway teow is easily available at most food centres in Singapore , such as at the Maxwell Road Hawker Centre , and it's also a signature dish at the Princess Terrace Café .	チャー・クウェイ・ティオは、「マックスウェル・ロード・ホーカーセンター」などのフードセンターで簡単に見つけることができます。 また、「プリンセス・テラス・カフェ」の名物料理でもあります。