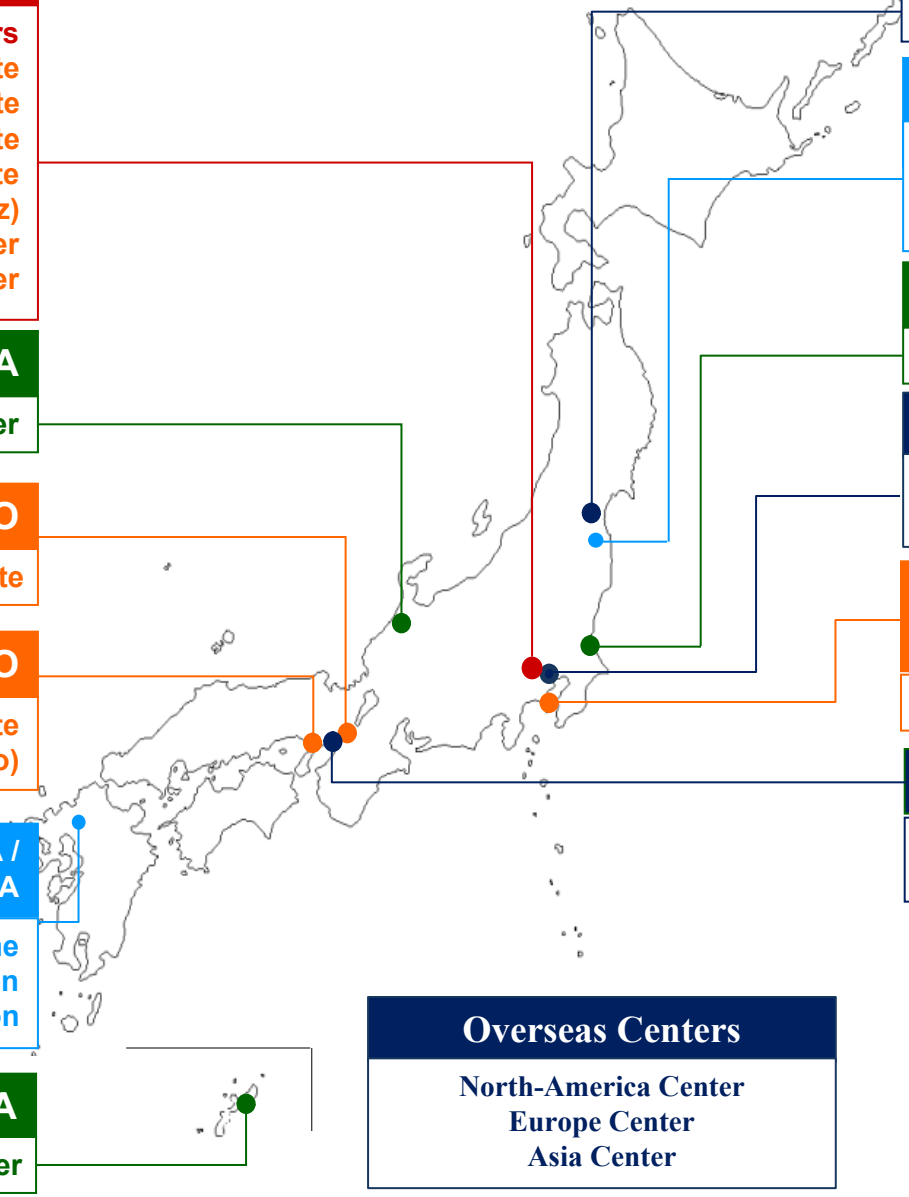
National Institute of Information and Communications Technology

Machine Translation of VoiceTra App

BarCamp 2016, Yangon, Myanmar

Ye Kyaw Thu,
Researcher,
Multilingual Translation Lab.,
Advanced Speech Translation Research and Development Promotion Centre,
NICT, Kyoto, Japan

Location of NICT Facilities



Koganei City, TOKYO
Headquarters
Photonic Network Res. Institute
Network Security Res. Institute
Applied Electromag. Res. Institute
Advanced ICT Res. Institute (Quantum, Terahertz)
Terahertz Technology Res. Center
Cybersecurity Research Center

Noumi City, ISHIKAWA
Hokuriku StarBED Techn. Center

Keihanna Region, KYOTO
Universal Comm. Res. Institute

Kobe City, HYOGO
Advanced ICT Res. Institute (Bio, Nano)

Saga City, SAGA / Itoshima-City, FUKUOKA
Hagane-yama LF Standard Time and Frequency Transmission Station

Onna Village, OKINAWA
Okinawa Electromag. Techn. Center

Sendai City, MIYAGI
Resilient ICT Research Center

Tamura City,
Ohtakadoya-yama LF Standard Time and Frequency Transmission Station

Kashima City, IBARAKI
Kashima Space Techn. Center

Chiyoda Ward, TOKYO
Network Testbed R&D Promotion Center

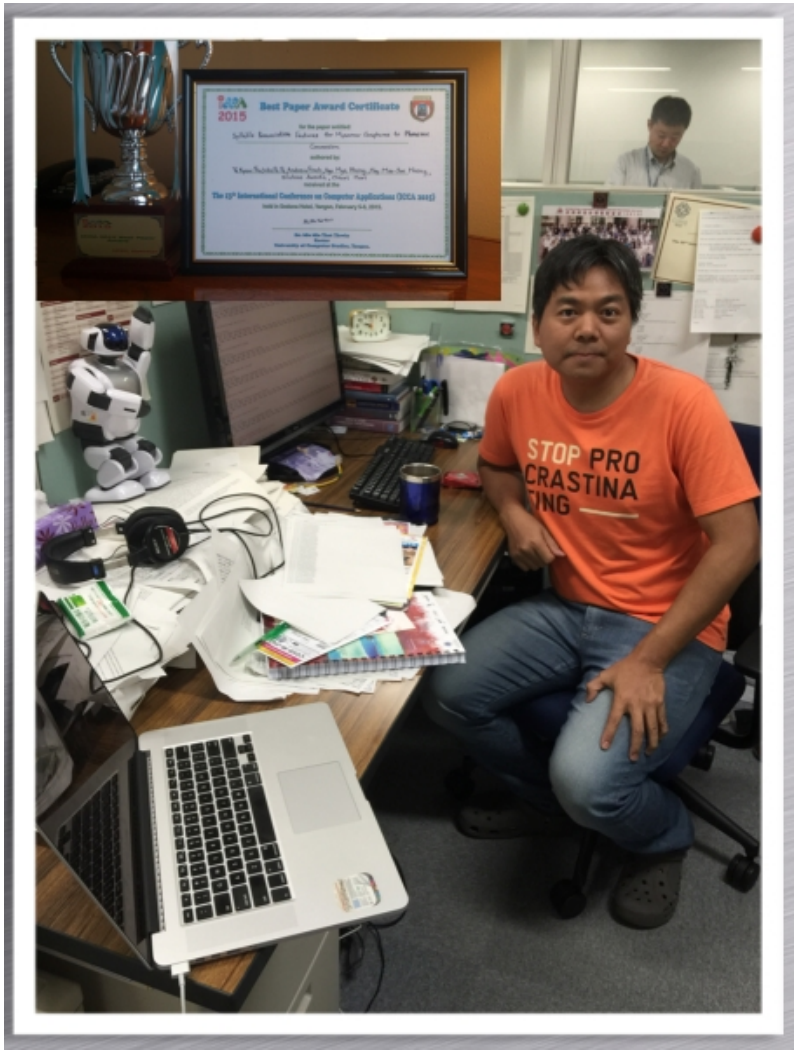
Yokosuka City, KANAGAWA
Wireless Network Res. Institute

Suita City, OSAKA
Center for Information and Neural Networks

Overseas Centers
North-America Center
Europe Center
Asia Center

- Headquarters
- Research Institute
- Promotion Center / Research Center
- Technology Center
- Standard Time and Frequency Transmission Station

>whoami



- A native of Myanmar
- Doctor's of Global Information and Telecommunication Studies from Waseda University
- Researcher at NICT
- Invited researcher at Waseda University

Home Page: https://sites.google.com/site/yekyawthunlp/

Table of Contents

MT History

“The use of computer software to translate text or speech from one language to another”

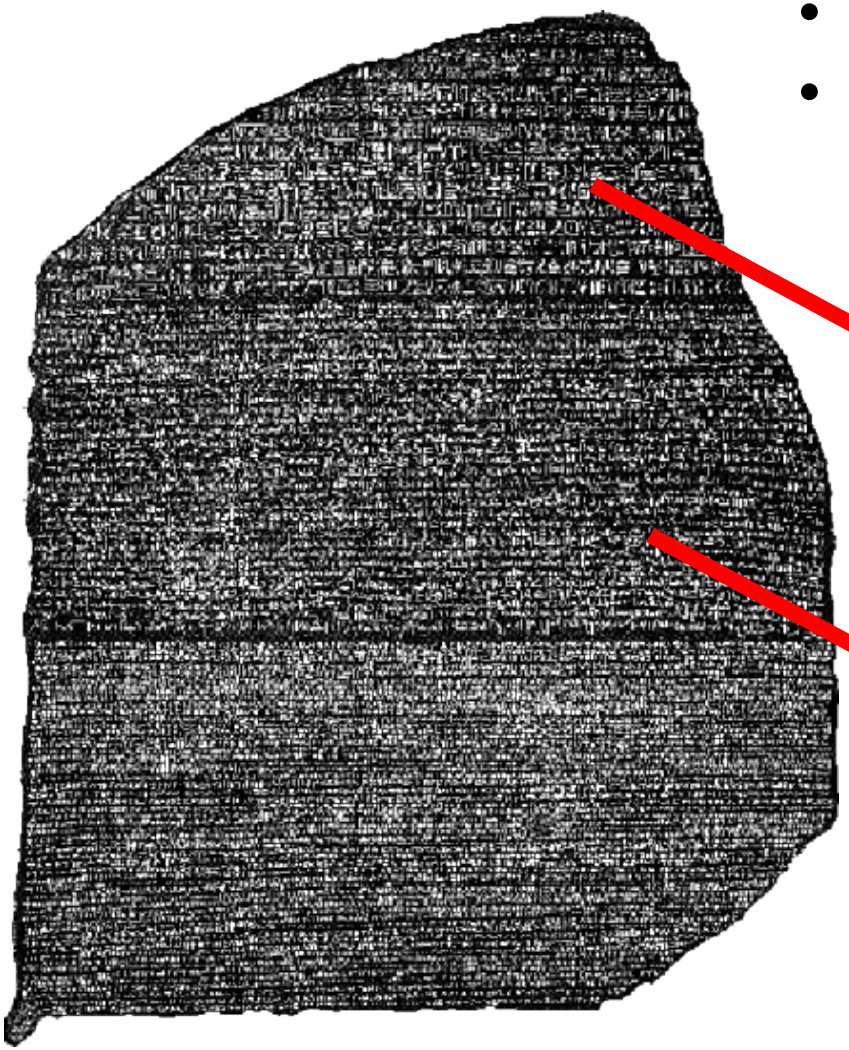
Wikipedia

MT History

Guess the Egyptian hieroglyphs from co-occurrence with Demotic words

(Jean-François Champollion,1822)

MT History



- Parallel text
- Egyptian writing system was a combination of phonetic and ideographic signs

MT History

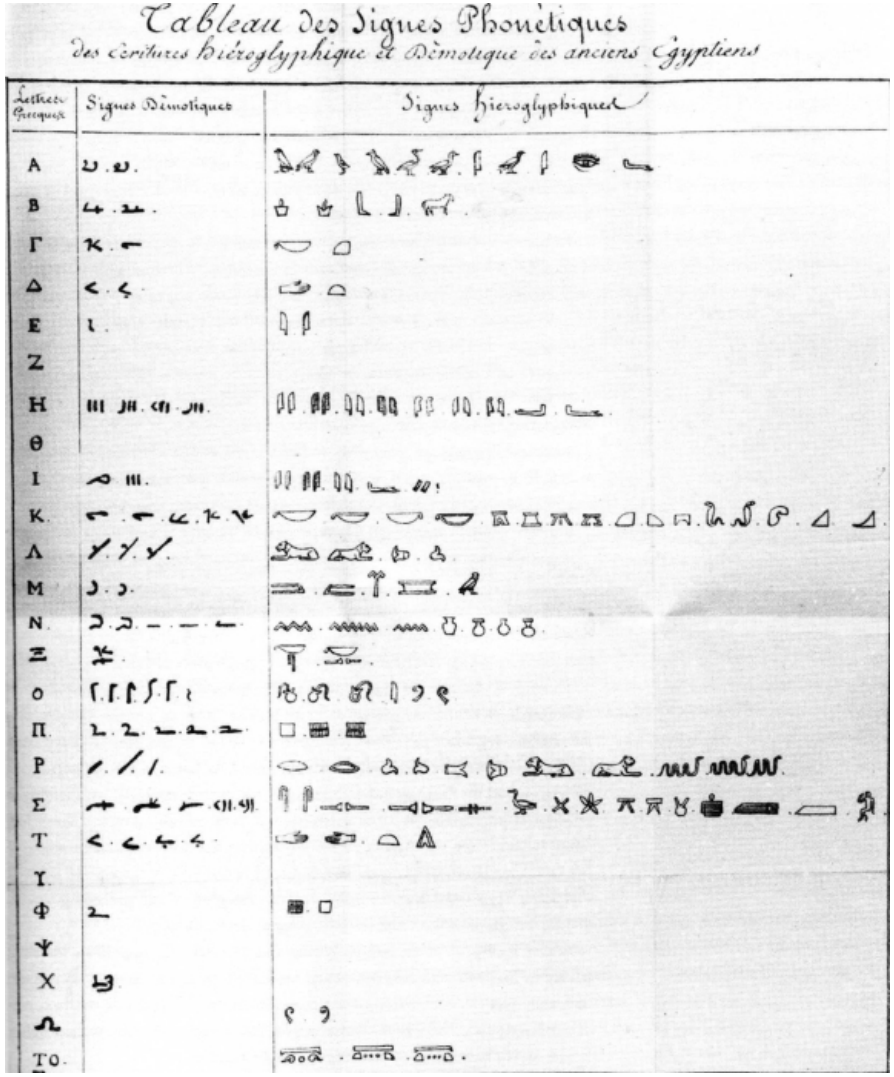
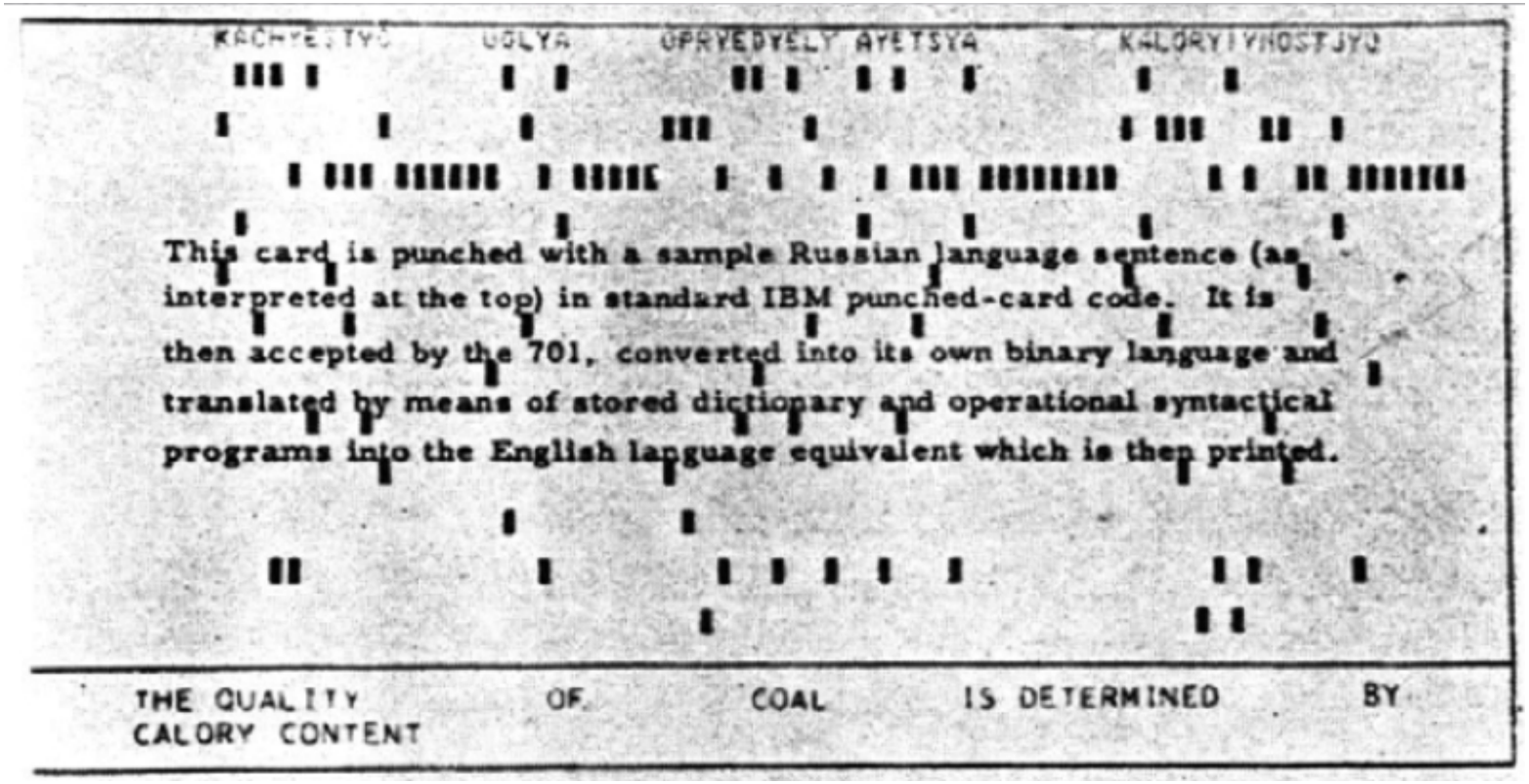


Table of hieroglyphic phonetic characters with their Demotic and Coptic equivalents (Champollion,1822)

MT History

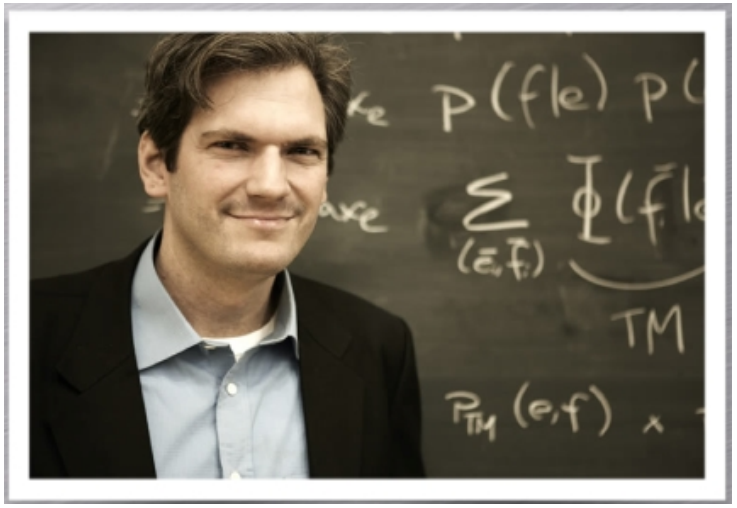
- Georgetown Experiment
- Collaboration between Gerogetown University and IBM in 1954
- Russia to English translation
- 250 words vocabulary
- Rule-based (6 rules of operational syntax)
- Over 60 Russian sentences translated automatically

MT History

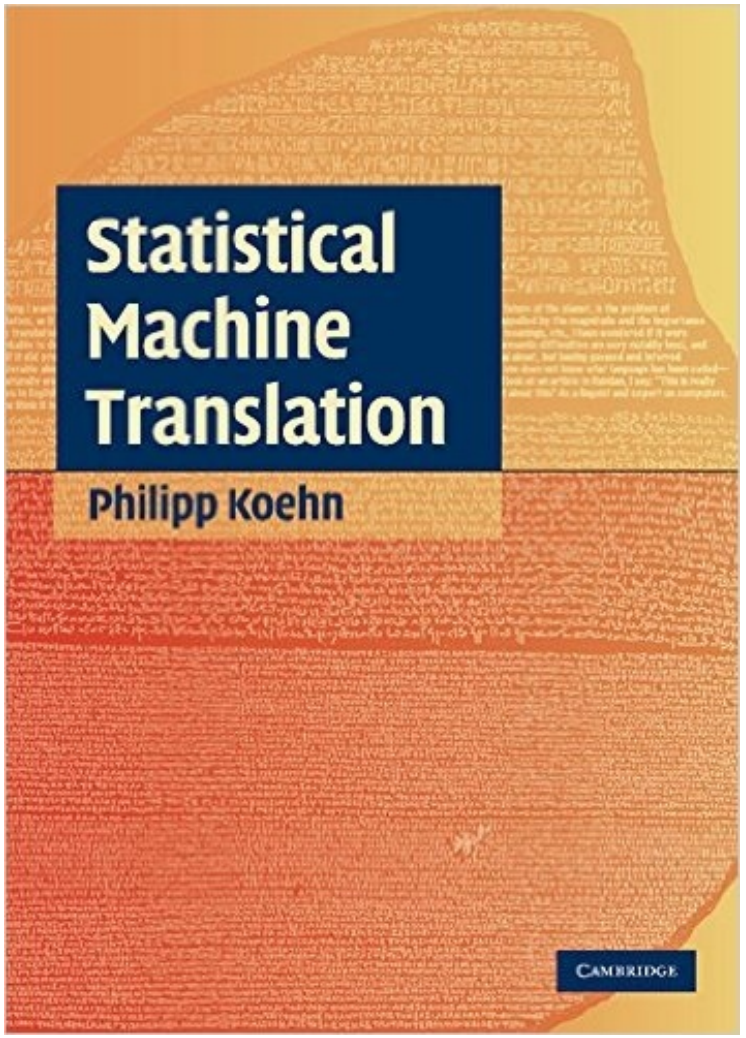


Russian → English Machine Translation

MT History



- Philipp Koehn
- One of the originator of phrase-based SMT



Corpus Preparation

こんにちは
မင်္ဂလာပါ

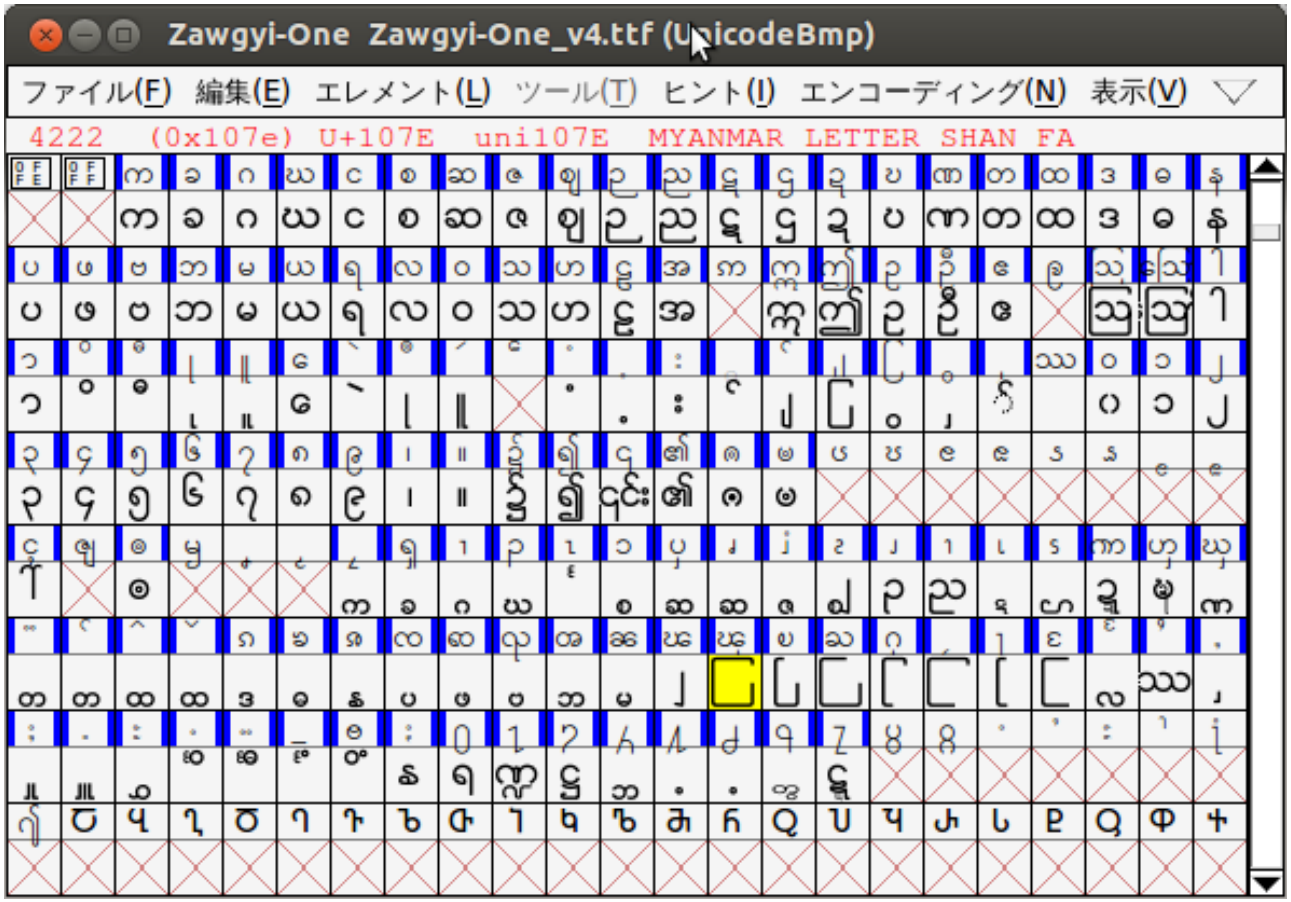
どうもありがとう
ကျေးဇူးတင်ပါတယ်။

バスで行けます
ဘက်စ်နဲ့သွားနိုင်တယ်

30分くらいです
မိနစ်သုံးဆယ်လောက်ပါ

- Parallel text for Japanese and Myanmar
- For Japanese to Myanmar SMT
- For Myanmar to Japanese SMT

Corpus Preparation



- Encoding issues (Myanmar3, Padauk or Zawgyi)
- We have to use UTF8 for SMT

Word Segmentation

prazer em conhecê-lo .
es freut mich , sie zu treffen .
pleased to meet you .
encantado de conocerle .
enchanté .
आपसे मिलकर खुशि हूँइ .
saya senang bertemu dengan anda .
piacere di conoscerla .
初めまして 。
만나 서 반갑 습니다 .
уулзахад таатай байна .
saya senang bertemu dengan anda .
တွေ့ ရတာ ဝမ်းသာ ပါတယ် ။
तपाईंलाई भेटेर खुसी लाग्यो ।
очень рад с вами познакомиться .
කරුණාකර ඔබ හමුවෙන්නට .

- How to break for Myanmar?!

Word Segmentation

```
Algorithm: Myanmar Syllable Breaking
Input: array A[1..n]
Output: array B[1..n*2]

j := 1;
char-type = NULL;
for   i=1,...,n   do
   char-type := Check-char-type(A[i]);
   if   char-type = 1   then
      if   (A[i-1] ≠ VIRAMA)   and
           (A[i+1] ≠ ASAT)   and
           (A[i+1] ≠  VIRAMA)   then
              B[j] := '_';
              B[j+1] := A[i];
           else
                B[j] := A[i];
        else if char-type = 2 or char-type = 3 or char-type = 4 then
                     B[j] := '_'
                     B[j+1] := A[i];
     else
                     B[j] := A[j];
j := j +1;
```

- the pseudocode for the Myanmar syllable segmentation algorithm

(I proposed at ICCA2013)

Word Segmentation

Unsegmented Input:		Segmented Output:
အားရှိတယ်။	⇒	အား_ရှိ_တယ်_။
အင်္ဂလိပ်	⇒	အင်္ဂလိပ်
မြန်မာကျောင်း	⇒	မြန်_မာ_ကျောင်း
ကုလသမဂ္ဂ	⇒	ကု_လ_သ_မဂ္ဂ

- Simple to implement, has high coverage, and is very accurate

Word Segmentation

$$P_{\boldsymbol{\lambda}}(\mathbf{Y}|\mathbf{W}) = \frac{1}{Z(\mathbf{W})} exp(\sum_{t=1}^{T}\sum_{k=1}^{|\boldsymbol{\lambda}|} \lambda_k f_k(y_{t-1}, \mathbf{W}, t))$$

- Using Conditional Random Fields (CRF)
  (J. Lafferty, 2001)

Character/syllable unigrams: $\{w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}\}$
Character/syllable bigrams: $\{(w_{t-1}, w_t), (w_t, w_{t+1})\}$
Character/syllable trigrams: $\{(w_{t-2}, w_{t-1}, w_t), (w_{t-1}, w_t, w_{t+1}), (w_t, w_{t+1}, w_{t+2})\}$

- For Myanmar (Win Pa Pa et. al, ICGEC2015)

Word Segmentation

Tag number	Tag	Position
1	<	The first syllable/character in a word
2	>	The second last syllable/character in a word
3	+	Represents both < and >
4	-	Others
5	|	Final syllable/character in a word

- Lists of segmentation tags (labeling)

Word Segmentation

Number of tags	Tag set
2	- |
3	< - |
4	< > - |
5	< > + - |

- Four tag sets used for segmentation

Word Segmentation

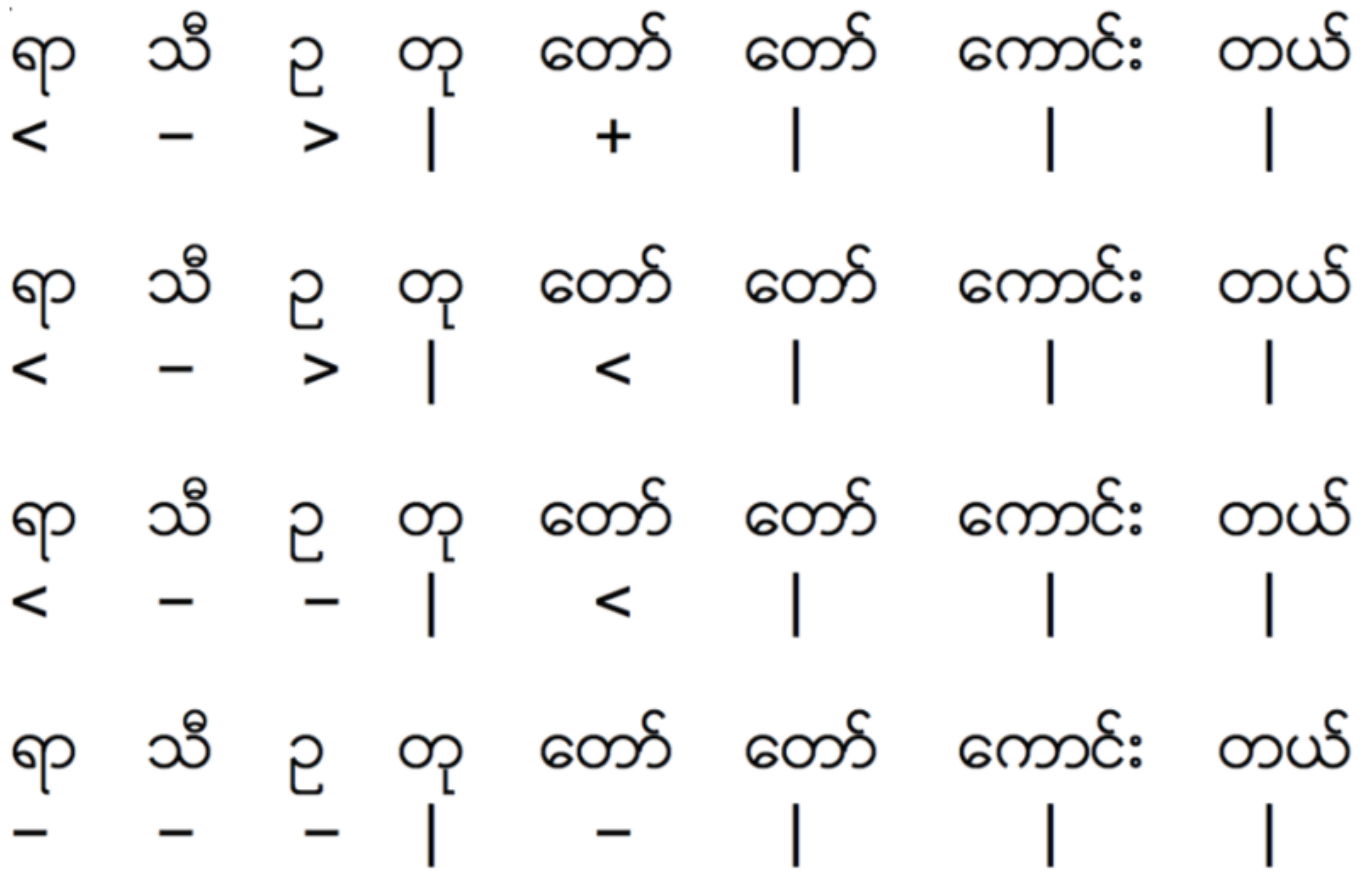


Fig. Syllable tagging with different tag sets

Word Segmentation

$$\text{F-score} = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

$$Precision = \frac{\#of\ correct\ tokens}{\#of\ tokens\ in\ test\ corpus}$$

$$Recall = \frac{\#of\ correct\ tokens}{\#of\ tokens\ in\ system\ output}$$

Fig. Evaluation with Precision, Recall and F-Score

Word Segmentation

Tagging Method	Character			Syllable		
	Precision	Recall	F-Score	Precision	Recall	F-Score
2 Tags	0.9695	0.9679	0.9687	0.9698	0.9683	0.9690
	±0.0040	±0.0056	±0.0046	±0.0035	±0.0048	±0.0040
3 Tags	0.9693	0.9686	0.9689	0.9703	0.9681	0.9692
	±0.0038	±0.0055	±0.0044	±0.0034	±0.0048	±0.0039
4 Tags	0.9694	0.9692	0.9693	0.9702	0.9676	0.9689
	±0.0038	±0.0053	±0.0043	±0.0034	±0.0048	±0.0040
5 Tags	0.9693	0.9692	0.9692	0.9703	0.9672	0.9687
	±0.0038	±0.0053	±0.0043	±0.0034	±0.0048	±0.0039

Fig. Word segmentation performance (with standard error) using different tag sets with CRF models

Word Segmentation

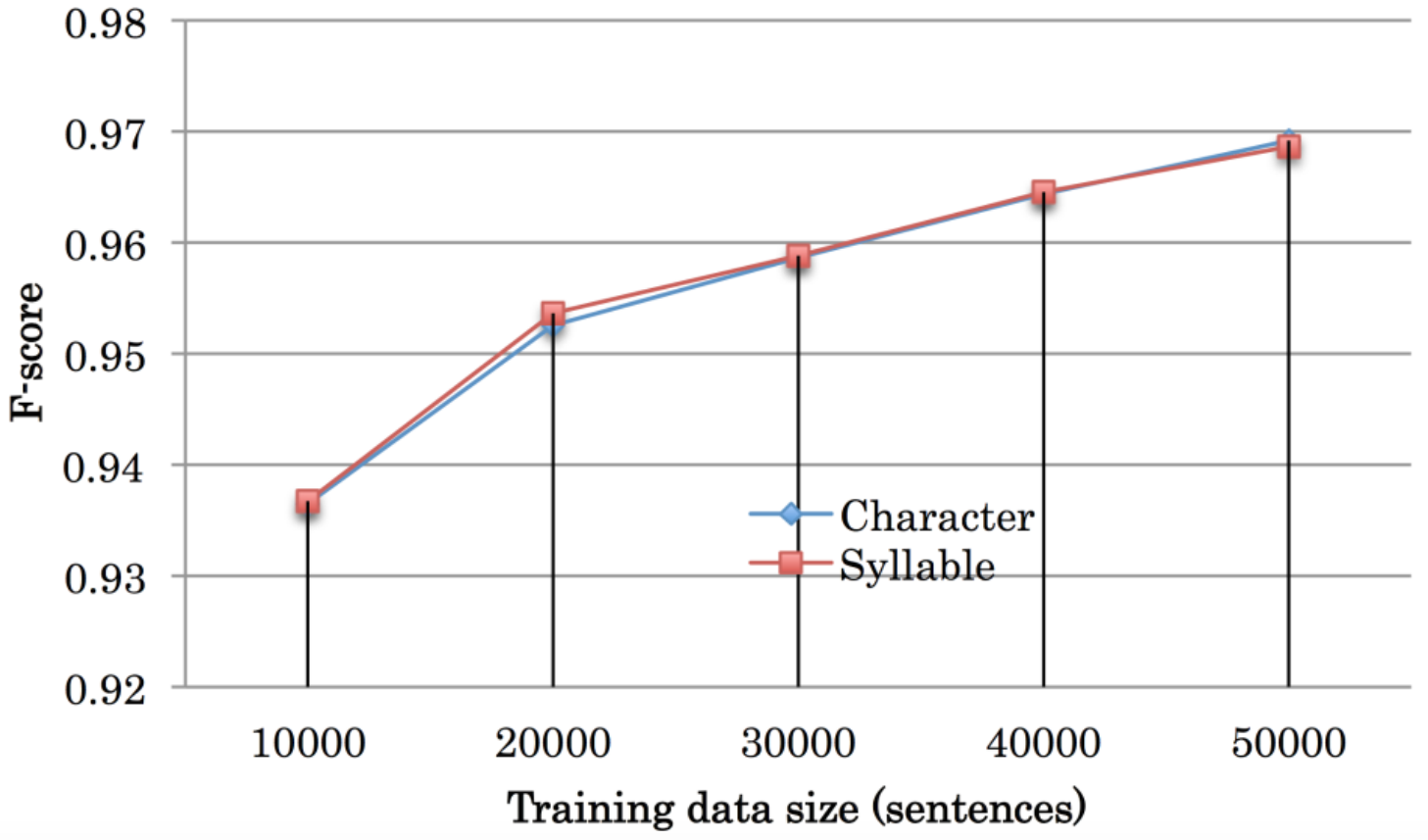


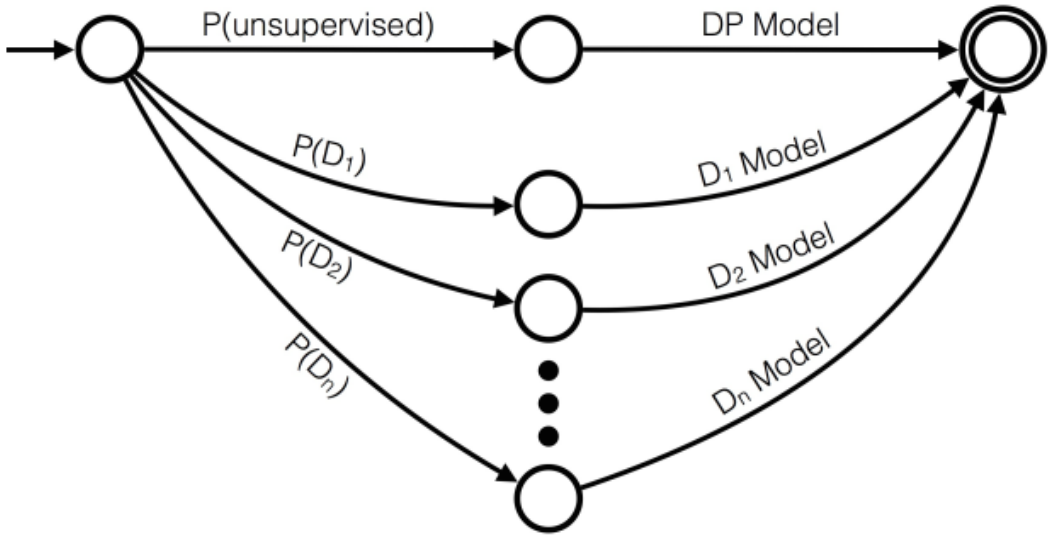
Fig. F-scores from training with CRF models on varying data set size

Word Segmentation

Reference	Output	Error Syllable	Percentage
|	+	ပေး,ပါ,တာ,သွား,ရ,ထား,ဘယ်,နေ,ပြန်,တစ်,တွေ,ရှိ,လိုက်,နား,တ,ဒီ,က	19.59
+	|	ပေး,သွား,ရှိ,တာ,ခေါ,တစ်,နစ်,တွေ,လာ,မှာ,လုပ်,ပြန်,ဆယ်,ပါ,ထား,တို,နေ	16.78
<	|	သွား,လို,ရှိ,ရ,ပေး,ဖြစ်,ထင်,နှစ်,လုပ်,တစ်,ထား	11.44
>	+	ပါ,ရ,မ,နာ,အ,နေ,စ,မှာ	9.49
|	<	ရ,ပါ,ချင်,နေ,လို့,ဖြစ်,ပေး	8.54
+	>	မ,ပါ,သ,စ,အ	8.29
<	+	အ,မ,ဘယ်,ဆောင်,နေ,ဒီ,ကြ	7.96
-	<	ရ,ပါ,နိုင်,ပေး	6.18
+	<	အ,ဘယ်,မ,နေ,နည်း	5.66
<	-	ပါ,နိုင်,ရ	3.38
>	|	လောက်,ရွက်,ရာ,တာ,တွေ,နား	1.62
|	>	လောက်,ပါ,နည်း	0.95

Fig. Statistics on the most frequent 300 labeling errors from 10 experiments for syllable tagging together with all the associated syllables

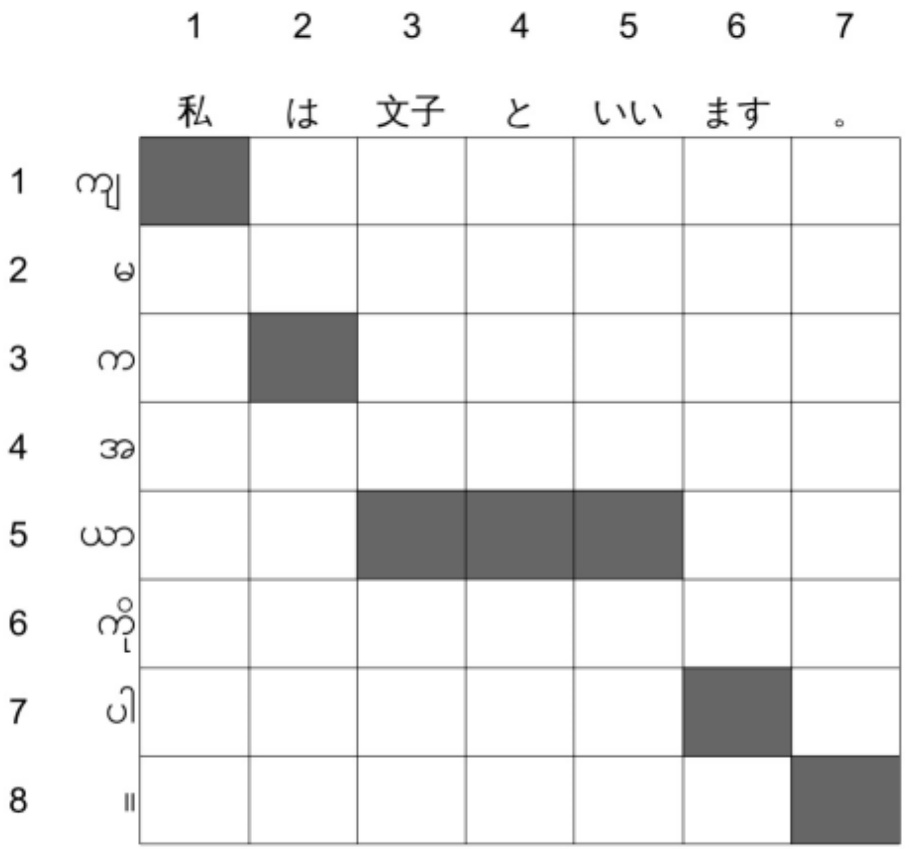
Semi-unsupervised & Semi-supervised (Examples relating to Myanmar language)



- Generative process with multiple dictionaries competing to generate the data alongside an unsupervised Dirichlet process model
- Refer to: “Integrating Dictionaries into an Unsupervised Model for Myanmar Word Segmentation”, (Ye et. al, COLING 2014)

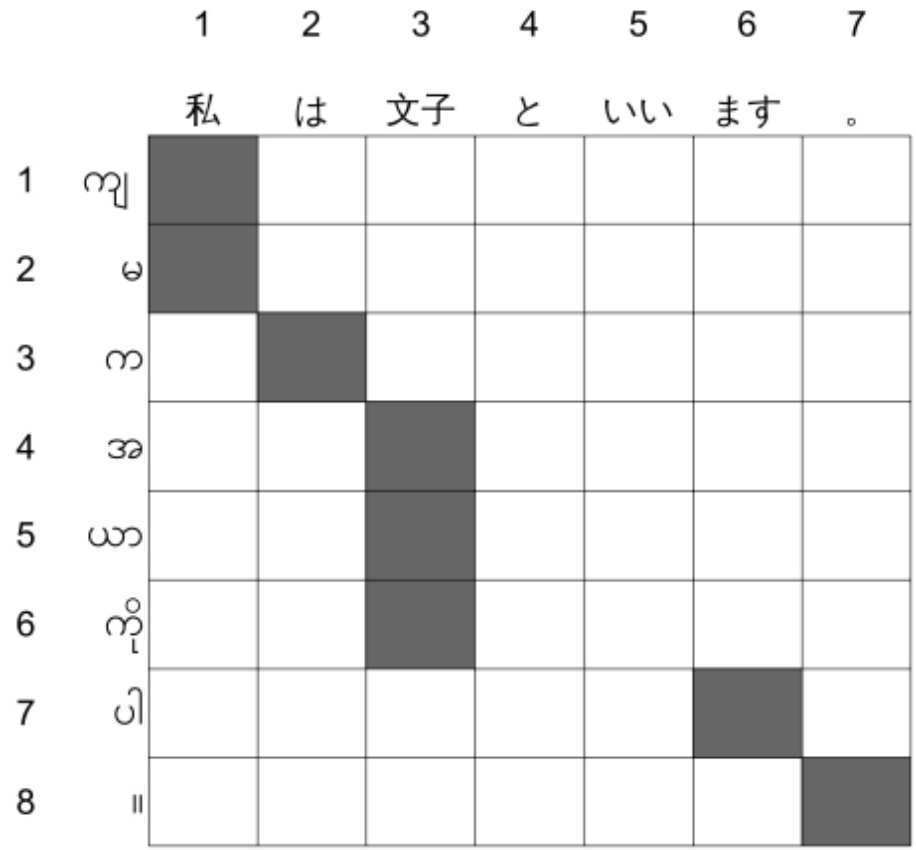
Alignment

- Example alignment for ja-my



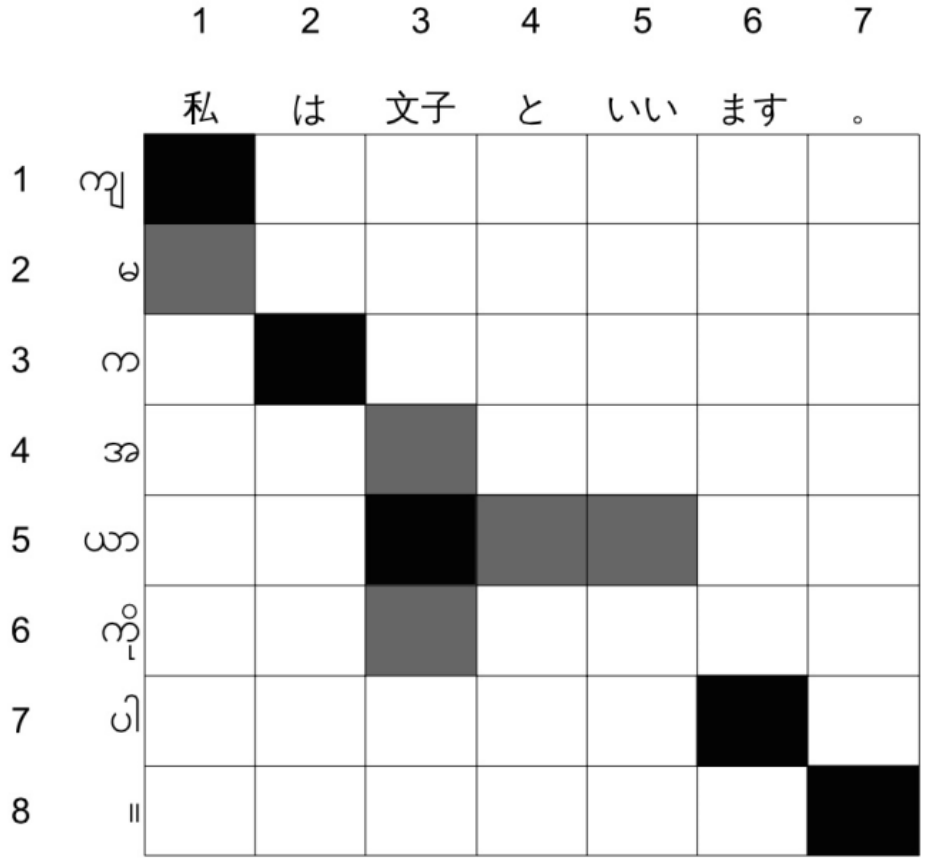
Alignment

- Example alignment for my-ja



Alignment

- Example alignment for ja-my, my-ja combination of GIZA++



Alignment

```
# Sentence pair (14) source length 6 target length 6 alignment score : 1.03383e-08
ဓာတ်ဆီ ဆိုင် က ဘယ်မှာ လဲ ။
NULL ({ }) where ({ 4 5 }) is ({ 3 }) the ({ }) gas ({ 1 2 }) station ({ }) ? ({ 6 })

# Sentence pair (15) source length 6 target length 6 alignment score : 2.59128e-07
သူ က ဘဏ် စာရေး ပါ ။
NULL ({ }) he ({ 1 }) is ({ 2 }) a ({ }) bank ({ 3 }) clerk ({ 4 5 }) . ({ 6 })

# Sentence pair (16) source length 6 target length 8 alignment score : 7.20491e-12
တယ်လီဖုန်း ကတ် က သုံး ရတာ အဆင်ပြေ တယ် ။
NULL ({ }) a ({ }) telephone ({ 1 }) card ({ 2 }) is ({ 3 }) handy ({ 4 5 6 7 }) . ({ 8 })
```

Fig. Inside my-en.A3.final.gz

- GIZA ++

(Franz Josef Och, Hermann Ney, 2003)

http://www.statmt.org/moses/giza/GIZA++.html

Phrase Table

```
" congratulation " from ||| ဂုဏ်ပြု ||| 0.0073484 2.69058e-11 0.073484 1 ||| 1-0 ||| 30 3 1 |||
" congratulation " from ||| ဂုဏ်ပြု ပါတယ် ||| 0.00165753 2.69058e-11 0.073484 0.0080614 ||| 1-0 ||| 133 3 1 |||
" congratulation " from ||| ဂုဏ်ပြု ပါတယ် လို့ ||| 0.036742 2.69058e-11 0.073484 9.19274e-05 ||| 1-0 ||| 6 3 1 |||
" congratulation " ||| ဂုဏ်ပြု ||| 0.0073484 1.66363e-08 0.073484 1 ||| 1-0 ||| 30 3 1 |||
" congratulation " ||| ဂုဏ်ပြု ပါတယ် ||| 0.00165753 1.66363e-08 0.073484 0.0080614 ||| 1-0 ||| 133 3 1 |||
" congratulation " ||| ဂုဏ်ပြု ပါတယ် လို့ ||| 0.036742 1.66363e-08 0.073484 9.19274e-05 ||| 1-0 ||| 6 3 1 |||
" congratulation ||| ဂုဏ်ပြု ||| 0.0073484 7.53522e-06 0.073484 1 ||| 1-0 ||| 30 3 1 |||
" congratulation ||| ဂုဏ်ပြု ပါတယ် ||| 0.00165753 7.53522e-06 0.073484 0.0080614 ||| 1-0 ||| 133 3 1 |||
" congratulation ||| ဂုဏ်ပြု ပါတယ် လို့ ||| 0.036742 7.53522e-06 0.073484 9.19274e-05 ||| 1-0 ||| 6 3 1 |||
```

Fig. Inside phrase-table.1.gz

- A big dictionary for Machine Translation
- You can also combine more than one phrase tables (several techniques)

Language Model

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

- Joint probability of words in a sentence

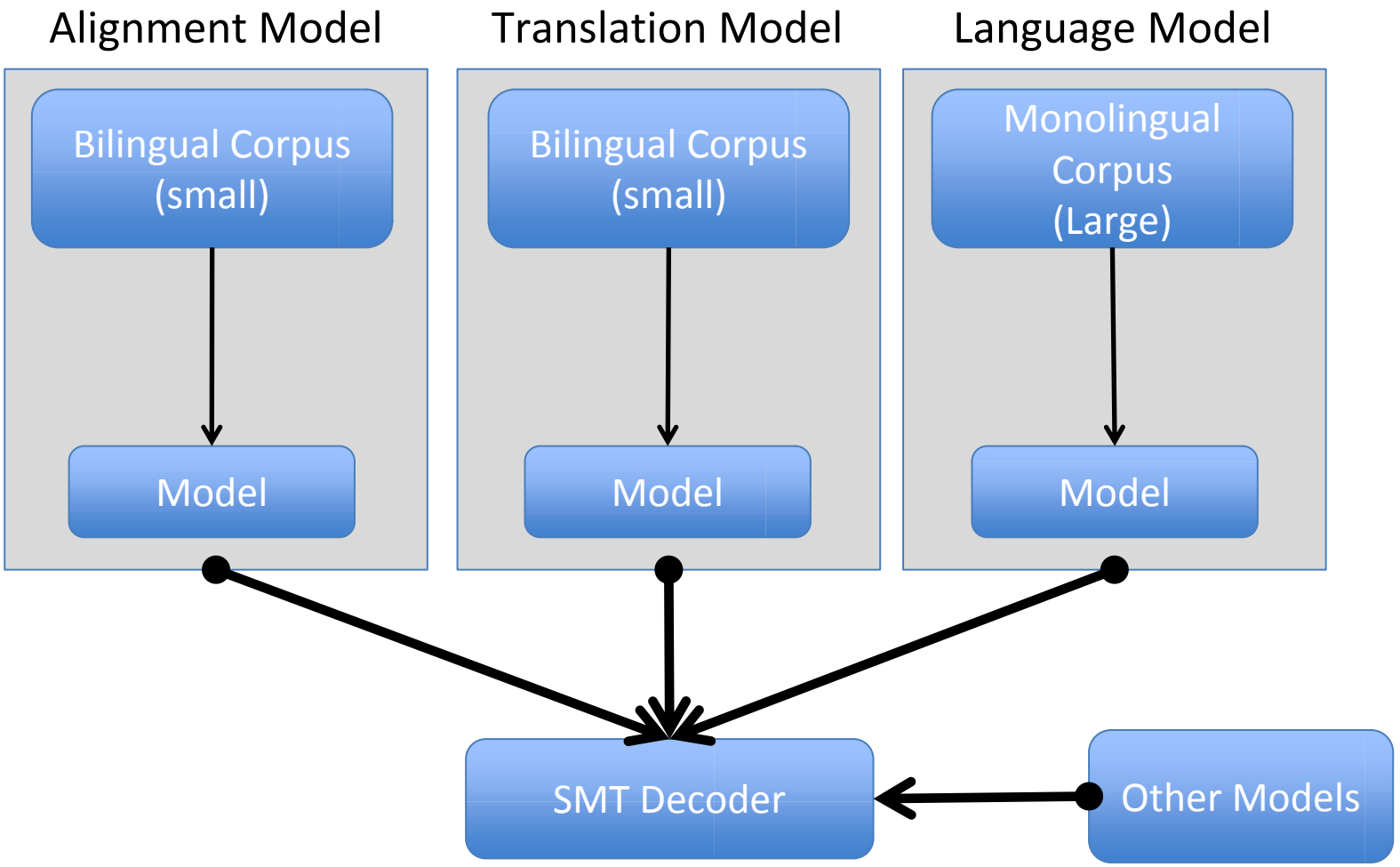
P(မင်းကဘယ်သူလဲ။)= P(မင်း) x P(က|မင်း) x P(ဘယ်သူ|မင်း က)
x P(လဲ|မင်း က ဘယ်သူ) x P(။|မင်း က ဘယ်သူ လဲ)

Language Model

```
-6.306803       ကိုယ့်စရိတ် -0.119332
-6.306803       ကိုယ့်စား  -0.119332
-6.306803       ကိုယ့်စားရိတ်       -0.119332
-5.742795       ကိုယ့်စိတ်  -0.119332
-6.306803       ကိုယ့်စိတ်ကိုယ်       -0.119332
-6.306803       ကိုယ့်ညီ     -0.119332
-6.306803       ကိုယ့်တာဝန်       -0.119332
-6.306803       ကိုယ့်တာဝန်ကိုယ်      -0.119332
-6.306803       ကိုယ့်တိုင်  -0.119332
-5.742795       ကိုယ့်ထက်  -0.1902117
-6.306803       ကိုယ့်ထက်ကြီး       -0.119332
-6.306803       ကိုယ့်ထက်ရာ       -0.119332
-6.306803       ကိုယ့်ထက်ရာထူး       -0.119332
-6.306803       ကိုယ့်ထမင်း       -0.119332
-6.306803       ကိုယ့်ဒုက္ခ  -0.119332
-6.306803       ကိုယ့်ဒူး    -0.119332
-6.306803       ကိုယ့်ဒူးခေါင်းကိုယ်  -0.119332
```

Fig. Inside language model file

Overall SMT



Moses Toolkit: http://www.statmt.org/moses/

Evaluation

- BLEU Score
  (de fecto standard scoring method)
- RIBES Score
  (for long distance language pair)

A Large Scale Study for Myanmar

- Basic Travel Expressions Corpus (BTEC)

- Languages:

  Arabic (ar), Chinese (zh), English (en), German (de), Hindi (hi), In- donesian (id), Italian (it), Japanese (ja), Ko- rean (ko), Malaysian (ms), Mongolian (mn), Myanmar (my), Nepali (ne), Portuguese (br), Russian (ru), Sinhala (si), Spanish (es), Tagalog (tl), Thai (th), Turkish (tl) and Vietnamese (vi)

- Corpus:

  457,249 sentences were used for training

  5,000 sentences for development and

  3,000 sentences for evaluation.

A Large Scale Study for Myanmar

- Phrase Based SMT (PBSMT)

  (Philipp Koehn, 2003)
- Hierarchical Phrase Based SMT (HPBSMT)

  (David Chiang, 2007)
- Operational Sequence Model (OSM)

  (Nadir Durrani, 2015)

A Large Scale Study for Myanmar (Corpus Statistics)

- We used twenty languages from the multilingual Basic Travel Expressions Corpus (BTEC)

Arabic (ar), Chinese (zh), English (en), German (de), Hindi (hi), Indonesian (id), Italian (it), Japanese (ja), Korean (ko), Malaysian (ms), Mongolian (mn), Myanmar (my), Nepali (ne), Portuguese (br), Russian (ru), Sinhala (si), Spanish (es), Tagalog (tl), Thai (th), Turkish (tl) and Vietnamese (vi)

A Large Scale Study for Myanmar (Corpus Statistics)

- Training: 457,249 Sentences
- Development: 5,000 Sentences
- Evaluation: 3,000 Sentences
- In all experiment: Myanmar language was segmented using rule based syllable segmentation, maximum matching and the CRF word segmentation methods

A Large Scale Study for Myanmar (Experimental Technology)

- Phrase-based (PBSMT)
  (Koehn and Haddow, 2009)
- Hierarchical phrase-based (HPBSMT)
  (Chiang, 2007)
- Operation sequence Model (OSM)
  (Durrani et al., 2015)

  Moses toolkit, GIZA++ (Och and Ney, 2000), grow-diag-final-and heuristic (Koehn et al., 2003), msd-bidirctional-fe (Tillmann, 2004), SRILM (5-gram LM) with interpolated modified Kneser-Ney discounting (Stolcke, 2002; Chen and Goodman, 1996), Minimum error rate training (MERT) (Och, 2003) for tuning, Moses decoder (ver. 2.1) (Koehn and Haddow, 2009)
- Evaluation with BLEU (Papineni et al., 2001) and RIBES scores (Isozaki et al., 2010)

(BLEU/RIBES * from my)

Src-Trg	Syllable			Word (Max-Match)			Word (CRF)		
	PBSMT	HPBSMT	OSM	PBSMT	HPBSMT	OSM	PBSMT	HPBSMT	OSM
my-ar	20.60 (0.63)	27.90 (0.68)	20.10 (0.61)	26.87 (0.67)	32.55 (0.71)	26.41 (0.67)	32.00 (0.70)	**34.16** (**0.72**)	32.09 (0.70)
my-br	29.21 (0.73)	37.88 (0.79)	29.17 (0.74)	35.57 (0.78)	40.56 (0.81)	35.80 (0.78)	39.12 (0.80)	**41.52** (**0.82**)	39.45 (0.80)
my-de	26.82 (0.74)	31.87 (0.78)	27.08 (0.74)	32.29 (0.77)	**35.90** (**0.79**)	32.78 (0.77)	34.98 (0.78)	35.80 (0.79)	35.14 (0.78)
my-en	33.14 (0.76)	42.76 (0.83)	33.31 (0.77)	40.28 (0.81)	45.86 (0.85)	39.83 (0.81)	43.82 (0.82)	**46.97** (**0.85**)	44.46 (0.83)
my-es	28.54 (0.72)	39.01 (0.79)	28.79 (0.73)	35.72 (0.78)	**42.69** (**0.82**)	35.35 (0.78)	39.49 (0.79)	42.02 (0.81)	40.08 (0.79)
my-hi	29.44 (0.70)	30.86 (0.71)	30.29 (0.70)	33.05 (0.71)	**33.87** (**0.73**)	33.45 (0.73)	33.31 (0.72)	33.71 (0.72)	33.74 (0.72)
my-id	29.63 (0.76)	39.25 (0.82)	29.86 (0.76)	36.28 (0.80)	42.04 (**0.84**)	36.55 (0.80)	41.38 (0.81)	**43.96** (0.83)	41.62 (0.82)
my-it	26.77 (0.70)	33.87 (0.74)	26.94 (0.70)	32.96 (0.75)	**37.57** (**0.78**)	33.39 (0.75)	35.69 (0.76)	37.41 (0.77)	35.96 (0.76)
my-ja	34.28 (0.79)	34.46 (0.79)	34.28 (0.79)	35.42 (0.79)	**35.77** (0.79)	35.50 (0.80)	35.36 (**0.80**)	35.36 (0.79)	35.57 (0.79)
my-ko	29.95 (0.74)	30.33 (0.74)	30.32 (0.74)	31.72 (0.75)	32.27 (0.76)	32.36 (0.75)	32.15 (**0.76**)	32.44 (0.76)	**33.03** (0.76)

(BLEU/RIBES * to my)

Src-Trg	Syllable			Word (Max-Match)			Word (CRF)		
	PBSMT	HPBSMT	OSM	PBSMT	HPBSMT	OSM	PBSMT	HPBSMT	OSM
ar-my	36.89 (0.82)	34.83 (0.81)	37.09 (0.82)	39.75 (0.83)	39.27 (0.83)	39.63 (0.83)	40.95 (0.83)	**41.63** (**0.84**)	41.06 (0.84)
br-my	38.11 (0.83)	38.66 (0.84)	38.40 (0.83)	40.26 (0.84)	41.62 (**0.85**)	40.38 (0.83)	41.85 (0.84)	**44.04** (0.84)	42.30 (0.84)
de-my	35.83 (0.82)	37.21 (0.83)	37.24 (0.83)	39.97 (0.84)	40.87 (0.85)	40.64 (0.84)	41.01 (0.84)	**42.16** (**0.85**)	41.10 (0.84)
en-my	39.46 (0.83)	41.22 (0.85)	40.95 (0.84)	42.96 (0.78)	44.15 (0.87)	42.40 (0.85)	44.87 (0.85)	**46.28** (**0.87**)	45.18 (0.86)
es-my	37.60 (0.82)	38.23 (0.84)	38.00 (0.82)	40.60 (0.83)	42.50 (0.85)	40.56 (0.83)	41.88 (0.84)	**44.03** (**0.86**)	41.69 (0.84)
hi-my	39.52 (0.84)	38.84 (0.84)	40.60 (0.85)	42.78 (0.86)	43.02 (0.85)	43.75 (0.86)	43.52 (0.86)	43.56 (0.85)	**43.80** (**0.86**)
id-my	38.39 (0.83)	38.05 (0.84)	39.60 (0.84)	41.14 (0.84)	42.21 (0.85)	42.00 (0.85)	43.59 (0.85)	**45.07** (**0.86**)	43.82 (0.85)
it-my	37.21 (0.82)	37.79 (0.84)	37.73 (0.83)	39.75 (0.83)	41.15 (**0.85**)	40.49 (0.83)	41.07 (0.83)	**41.57** (0.85)	41.50 (0.84)
ja-my	33.00 (0.82)	33.41 (0.82)	33.56 (0.82)	35.24 (0.82)	35.37 (0.82)	35.61 (0.82)	35.88 (0.82)	**36.57** (**0.82**)	35.91 (0.82)
ko-my	33.72 (0.83)	33.66 (0.83)	34.68 (0.83)	36.42 (0.83)	36.60 (0.83)	36.22 (0.83)	37.43 (0.83)	**37.94** (**0.84**)	37.14 (0.83)

A Large Scale Study for Myanmar (Discussion & Analysis)

- From the results, it is clear that the supervised CRF-based word segmentation scheme was by far the most effective
- From **my-***, a few cases 30% (Maximum Matching gave better results in terms of BLEU score)
- For ***-my**, the CRF approach dominated, outperforming the other word segmentation approaches for all language pairs

A Large Scale Study for Myanmar (Discussion & Analysis)

- The highest absolute BLEU scores were achieved on the Myanmar-English and English-Myanmar
- The reason for this is not because it is easy to translate between these two languages
- Most of the Myanmar part of the corpus was created by translating from the English sentences
- Another factor: may have contributed here is that the English data (along with the Japanese) has been subjected to the most checking and

A Large Scale Study for Myanmar (Discussion & Analysis)

- Our motivation for studying the application the OSM and HPBSMT techniques for Myanmar
- From the results, PBSMT approach was not the most effective approach (see Table 1, **my-***) and HPBSMT approach was the most effective (75% of the experiments)
- Again for ***-my**, HPBSMT approach was the most effective for 80% of the experiments and the OSM approach give the best score in 15% of the experiments

A Large Scale Study for Myanmar (Discussion & Analysis)

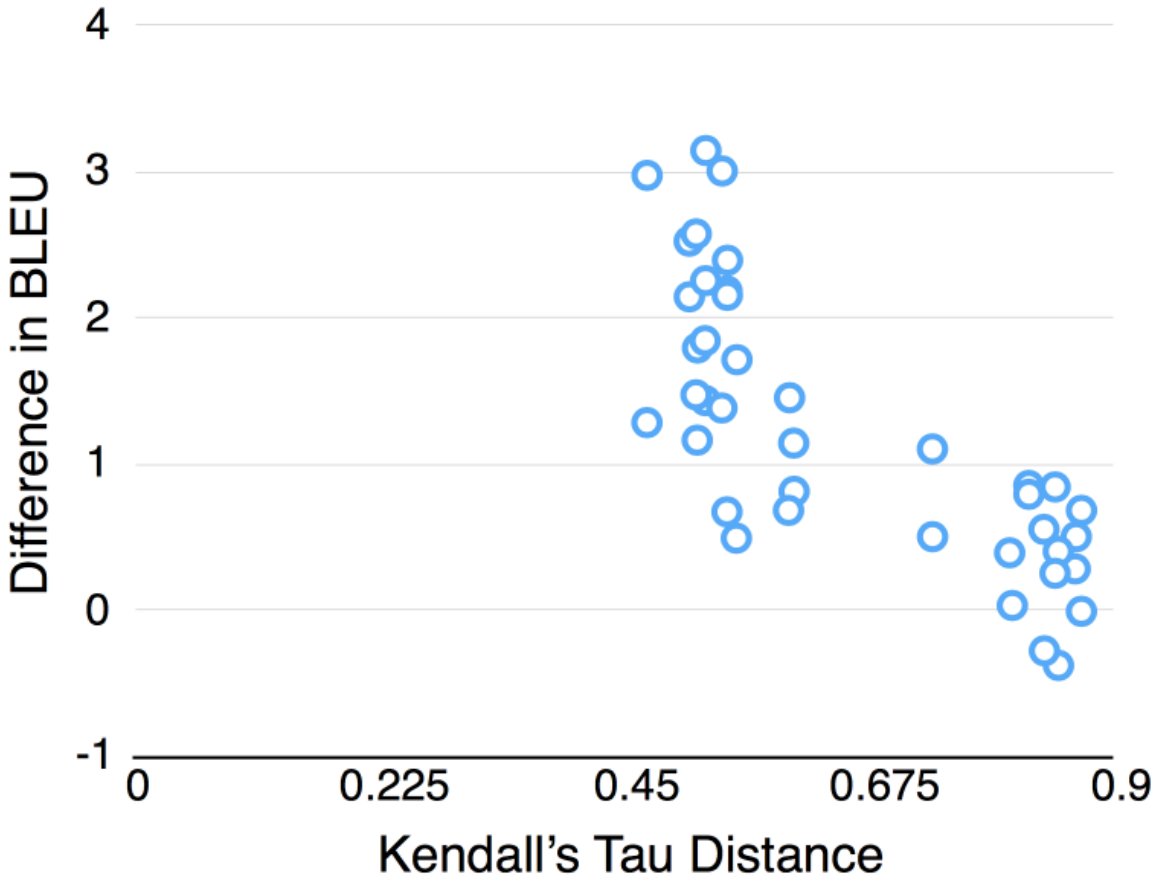


Figure. Plot of the Kendall's tau distance against BLEU difference

Conclusion

- Many things to do for Myanmar NLP
- If you want to try, download Moses and try it

  (Most of the tools are OpenSource)
- Note:

  Developing and doing research are different!
- Don’t forget! We need large parallel data

Thank You!

A Large Scale Study for Myanmar(Discussion & Analysis)

- We analyzed the results using Kendall's tau distance in order to guage the effect of re-ordering
- We calculated the Pearson product-moment correlation coefficient (PMCC) between the BLEU score and the Kendall's tau distance to assess the strength of the linear relationship

A Large Scale Study for Myanmar

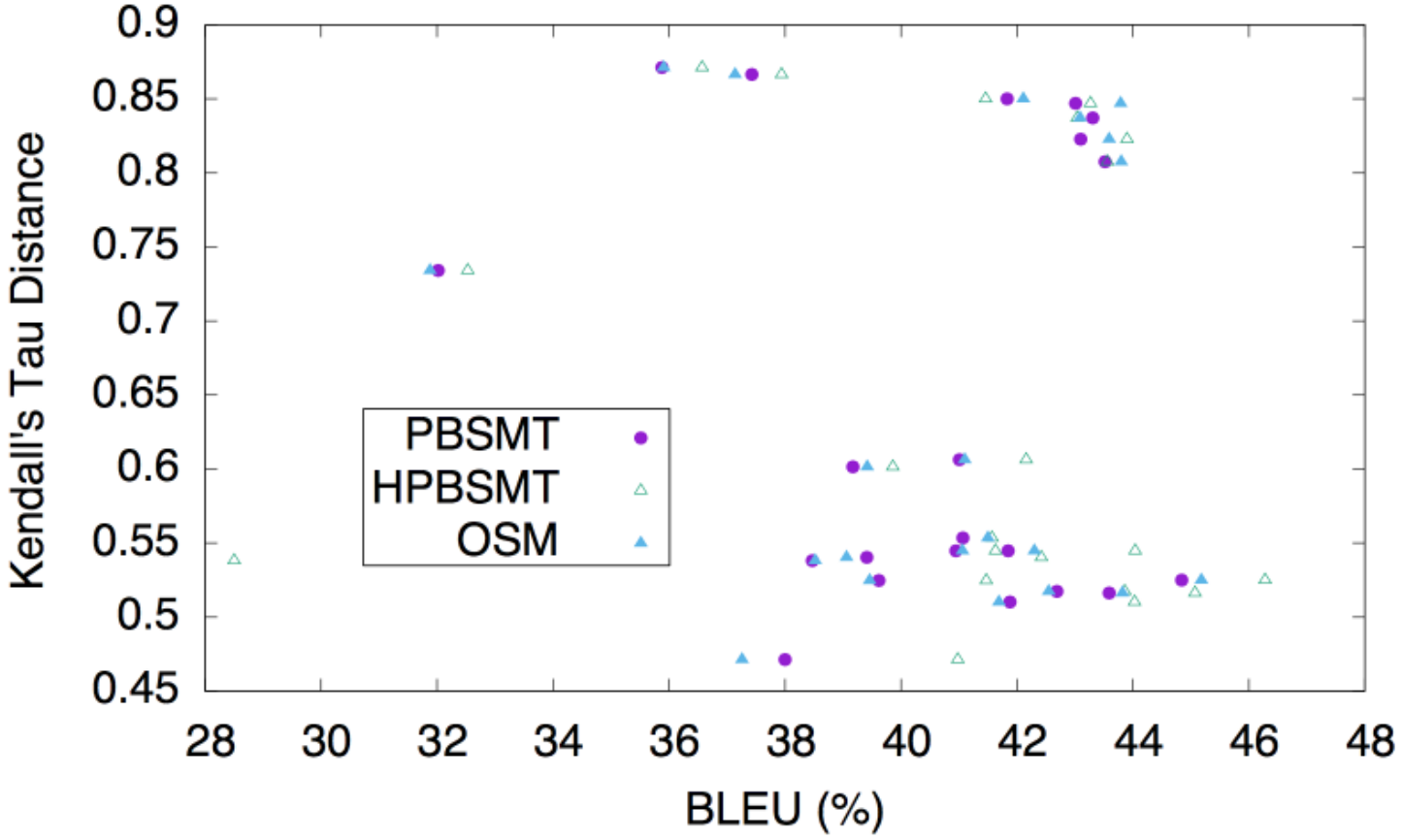


Fig. Plot of the Kendall's tau distance against BLEU