A Study of Machine Translation for Low-Resource Languages

Kumamoto University

- Sari Dewi Budiwati
- Supervisor: Prof. Masayoshi Aritsugi

- *Pre-defense, July 28th, 2020*

Outline

- Background
- Basic theory and related work
- Preliminary work
  - Single pivot in Kazakh to English (Kk-En)
  - Single and multiple pivots in Japanese to Indonesian (Ja-Id)
- Proposed approach
  - Word reordering in multiple pivots for Japanese to Indonesian (Ja-Id)
  - The comparison of phrase table combination for Kazakh to English (Kk-En) and Japanese to Indonesian (Ja-Id)
- Conclusion and future work

Background

- Machine Translation (MT) is a task of automatically translate a text from one natural language, i.e., English, to another language, i.e., Japan
- The state-of-the-art of MT
  - Statistical Machine Translation (SMT)
    - ➔ SMT is an approach that uses probabilistic models of faithfullness and fluency and then combining these models to choose the most probable translation (Jurafsky and Martin, 2009)
  - Neural Machine Translation (NMT)
    - ➔ NMT is based on neural network model that consist of encoder-decoders (Bahdanau et al., 2015)

Background

- Koehn and Knowles (Koehn and Knowles, 2017) compared two models, and stated that the NMT model still has to overcome various challenges
  - Performance of out-of-domain
  - Performance of low-resource conditions
- Low-resource conditions is a state where two language pairs have
  - Limited parallel corpora
  - Limited linguistic tools
- There are two strategies to achieve high-quality in low-resource (Trieu, 2017)
  - Building parallel corpora
  - Utilizing existing corpora

Background

- In this study, we explore pivot approaches in the SMT model for two language pairs:
  - Kazakh to English (Kk-En)
    - ➔ Available parallel corpora: 953,240
  - Japanese to Indonesian (Ja-Id)
    - ➔ Available parallel corpora: 1,468,155
- Objectives
  - To apply pivot approaches and examine issues in two low-resource language pairs
  - Proposed a technique that could improve the translation quality compare to the direct translation

Contribution

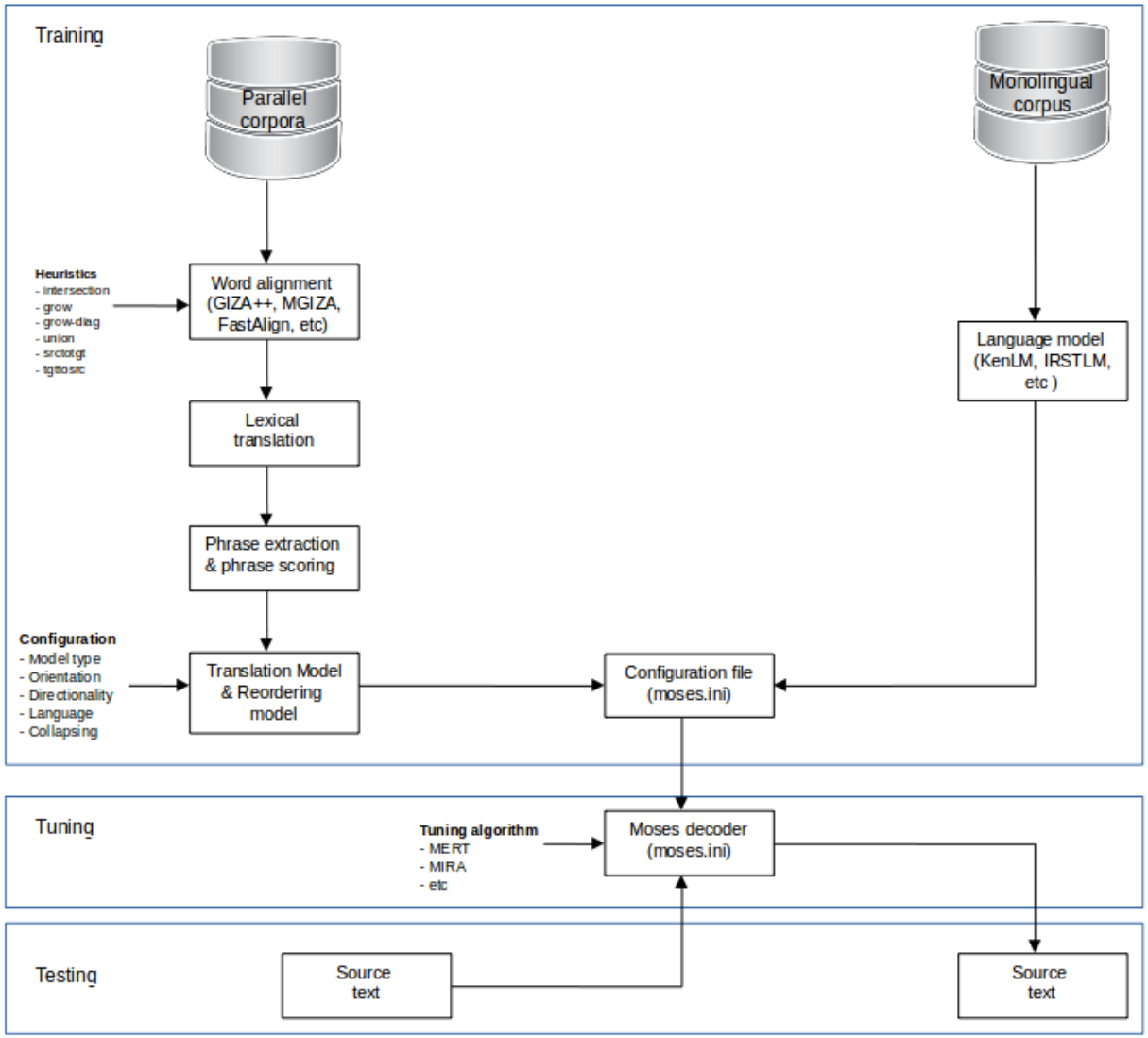
- In the single pivot, we proposed a phrase table combination based on different symmetrization technique.
  - These techniques come based on the fact that the pivot approach comprises three direct translations, viz., src-trg, src-pvt, and pvt-trg, that obtained different BLEU scores when different symmetrization employed.
  - Therefore, we did phrase table combinations based on the first and second highest BLEU scores of direct translation.
  - Our approach is competitive because it can improve the translation for Kk-En by more than 0.22 compared to the direct translation. Whereas the Ja-Id still obtained the best BLEU score by direct translation by more than 0.12 compared to our approach

Contribution

- In multiple pivots of Ja-Id, we proposed an extending src-pvt phrase table before the phrase table combination process.
  - These techniques arise based on our finding that the non-standard symmetrization has candidate phrase pair that could not be obtained by the standard one.
- In multiple pivots of Ja-Id, we employed *pre-ordering* of Japanese dataset to overcome the different word order between Japanese and Indonesian

Basic theory and related work

Statistical Machine Translation



Components of SMT
- Parallel corpora
- Monolingual corpora
- Translation Model (TM)
- Language Model
- Reordering Model
- Decoder

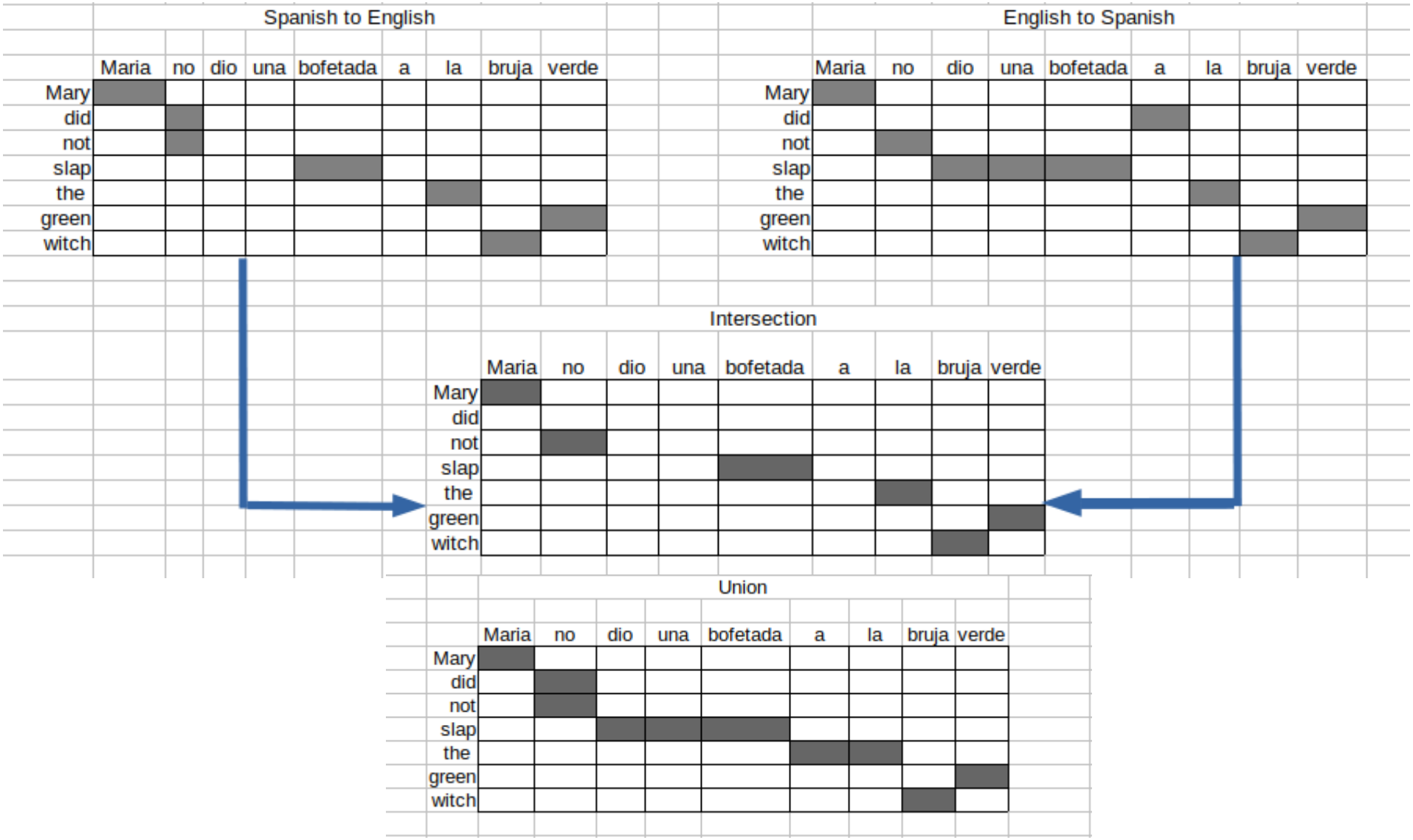
SMT process
1. Training
2. Tuning
3. Testing
4. Evaluation

Kumamoto University

Basic theory and related work

- In order to train the parallel corpora for phrase table, the SMT needs word alignment symmetrization, viz., intersection (I), union (U), grow, grow-diagonally, grow-diag-final, grow-final, source to target (srctotgt), target to source (tgttosrc)

Basic theory and related work

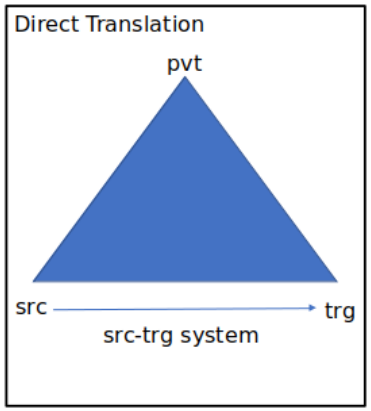


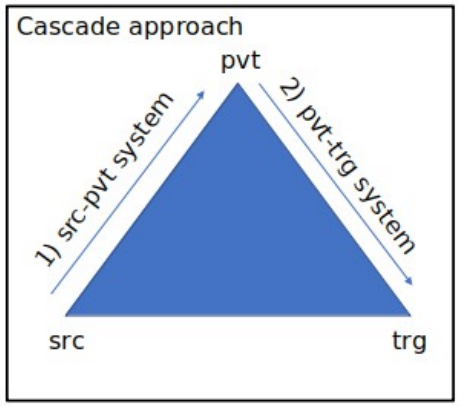
Word alignment symmetrization examples (Jurafsky and Martin, 2009)

Kumamoto University

Basic theory and related work

- Direct translation
  - src → trg
- Pivot approach
  - Pivot approach is a translation from a source language (src) to a target language (trg) through an intermediate pivot language (pvt) Paul et al. (2009).
  - Sentence translation (cascade)
    - ➔ src-pvt system
    - ➔ pvt-trg system

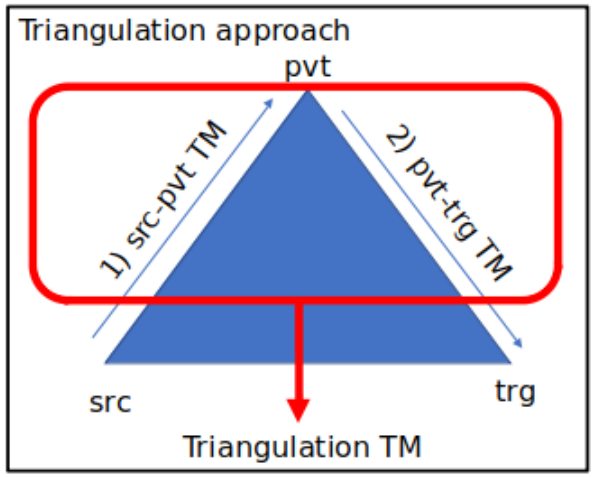


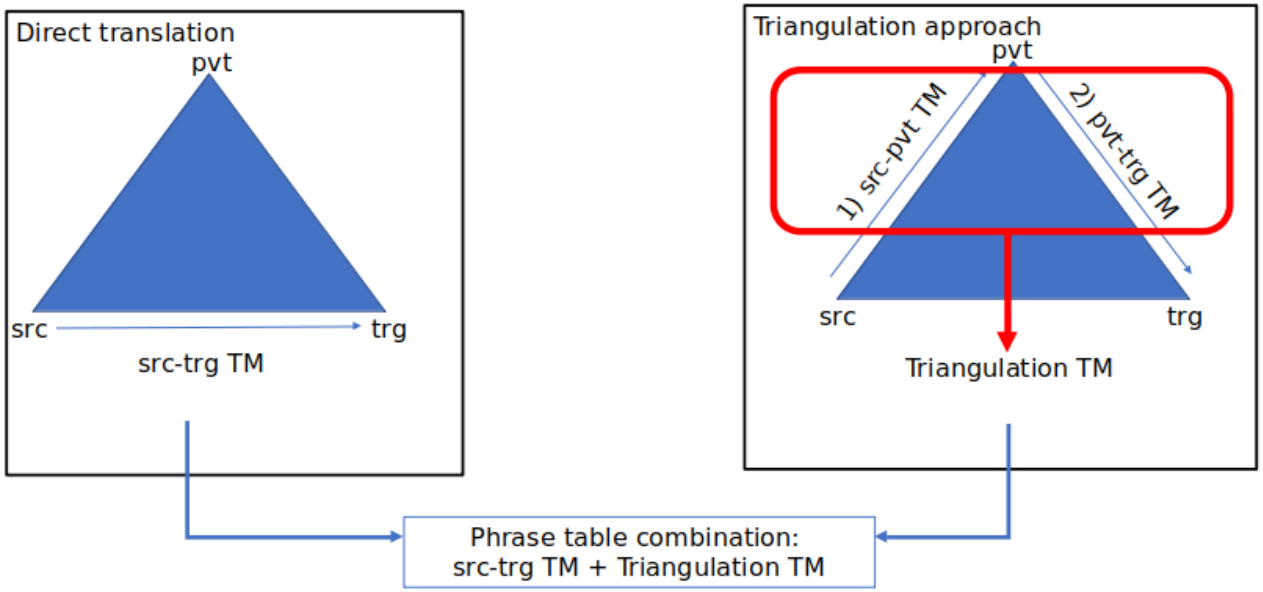


Kumamoto University

Basic theory and related work

- Pivot approach
  - Triangulation
    - → src-pvt TM
    - → pvt-trg TM
  - Phrase table combination





Kumamoto University

Basic theory and related work

- Kazakh to English (Kk-En) Machine Translation
  - The morphological segmentation has been used in SMT model (Assylbekov and Nurkas, 2014; Kartbayev, 2015b).
    - ➔ Kazakh is agglutinative language
    - ➔ Morphological segmentation is an approach that break words into morphemes
  - Shared task in Workshop Machine Translation (2019)
    - ➔ Most participants used NMT model with several approaches, i.e., back translation, transfer learning, multilingual transfer learning, sequence2sequence
    - ➔ The transfer learning is a similar technique as the pivot approach in the SMT model.
      - Use high-resource language pair to train the parent model.
      - The parent training data replaced with the training data of low-resource language pair

Kumamoto University

Basic theory and related work

- The WMT results denotes that the third language still needed to improve the translation quality of Kk-En, despite in NMT model.
- Our works in Kk-En
  - We proposed a phrase table combination that uses different symmetrization for Kk-En
  - We use different symmetrization because the translation quality of src-trg, src-pvt, and pvt-trg obtained different BLEU score when non-standard symmetrization employed

Basic theory and related work

- Japanese to Indonesia (Ja-Id) machine translation
  - Paul et al. (2009) and Adiputra and Arase (2017)
    - Focused on finding a technique that could improve the BLEU scores
    - Paul et al (2009) find that non-English could be used to improve the translation quality in pivot approach, i.e., cascade
    - Adiputra et al (2017) find that the SMT model outperformed the NMT model
  - Simon and Purwarianti (2013), Sulaeman and Purwarianti (2015)
    - Focused on finding a technique that could resolve the morphological issue, i.e., word order problem, incorrect defined phrase, and words with affixes,
    - Simon and Purwarianti (2013) proposed several techniques, i.e., using pos-tag, increasing the LM dataset, stemming for Indonesian dataset, removing Japanese particles, removing NE
    - Sulaeman and Purwarianti (2015) proposed several techniques, i.e., the pos-tag model, the hierarchical model, lemmatizer, and post-processing.
    - Both works did not show the generated text result for word ordering issue. Thus, it is hard to compare the pre-proposed and post-proposed of generated text

Basic theory and related work

- Several experimental results show that
  - The SMT model still a primary option, mainly using a pivot approach
  - It needs an additional technique to overcome the morphological issue
- Our work in Ja-Id
  - Employ pre-ordering technique in multiple pivots
  - We proposed two technique
    - ➔ The extended phrase table of src-pvt
    - ➔ Phrase table combination based on different symmetrization

Basic theory and related work

- Characteristics of our works
  - In the phrase table combination, most researchers use standard symmetrization, i.e., gdfand. In contrast, we use different symmetrization because the translation quality of src-trg, src-pvt, and pvt-trg obtained different BLEU scores when non- standard symmetrization employed.
  - In extended phrase table of src-pvt, we employ our finding, i.e., the unknown words of gdfand is available in tgttosrc phrase table. As a result, the phrase pair could replace the unknown word of gdfand.

Preliminary work

Single and multiple pivots in two low-resource languages

1. Single pivot in Kk <--> En
2. Single and multiple pivots in Ja <--> Id

Single and multiple pivots in two low-resource languages

- Single and multiple pivots in Kazakh to English (Kk-En)
  - Background
    - Our participation in Workshop on Machine Translation (WMT) 2019
    - Low-resource language: Kk ↔ En
    - Kk-En is the first language pair exploration in annual WMT. There is no experience system description from previous WMT.
    - Dataset domain: news
- Our works
  - We build two systems:
    - Baseline is a direct translation system
    - Interpolation is a pivot system:
  - Russian as a pivot language
  - We use Linear Interpolation/LI (phrase table combination) as a pivot approach. We use two LM orders, i.e., 3-gram and 5-gram.

Kumamoto University

Single and multiple pivots in two low-resource languages

BLEU scores result

- The interpolation system obtained high BLEU score compare to Baseline.
- The improvement for Kk-En is 0.1 and 0.5 for 3-gram and 5-gram
- The improvement for En-Kk is 0.1 for 3-gram and 5-gram

Table 3.2: BLEU-cased score results

Language Pair	3-gram LM	5-gram LM
KK-EN		
1. Baseline system	2.6	2.9
2. Interpolation system	2.7	3.4
EN-KK		
1. Baseline system	0.8	0.8
2. Interpolation system	0.9	0.9

Single and multiple pivots in two low-resource languages

- Perplexity score result
  - ➔ Lower perplexity score indicates better LM, while high perplexity scores indicates poor LM
  - ➔ The lowest perplexity obtained by 5-gram LM in Kk-En and En-Kk

Table 3.3: Perplexity results

Language pair	3-gram LM	5-gram LM
KK-EN		
1. Baseline system	- Incl OOVs: 829.59 - Excl OOVs: 77.79	- Incl OOVs: 617.36 - Excl OOVs: 45.51
2. Interpolation system	- Incl OOVs: 1034.50 - Excl OOVs: 94.72	- Incl OOVs: 762.79 - Excl OOVs: 50.93
EN-KK		
1. Baseline system	- Incl OOVs: 328.940 - Excl OOVs: 103.27	- Incl OOVs: 256.138 - Excl OOVs: 77.185
2. Interpolation system	- Incl OOVs: 256.13 - Excl OOVs: 79.34	- Incl OOVs: 276.85 - Excl OOVs: 85.40

Single and multiple pivots in two low-resource languages

Single and multiple pivots in Japanese to Indonesian (Ja-Id)

- We use four pivot languages: English, Myanmar, Malaysia, and Filipino, from ALT dataset (Riza et al., 2016)
- We build several systems, as follows:
  - Baseline
  - Single pivot
    - Cascade approach
    - Triangulation approach
    - Interpolation approach
  - Multiple pivots: combination of four phrase tables using Linear and Fillup Interpolation
    - All-LI: combine four phrase tables without src-trg by LI
    - All-FI: combine four phrase tables without src-trg by FI
    - Base-LI : Combine scr-trg and All-LI by LI
    - Base-FI : Combine src-trg and All-FI by FI
- Dataset type: Sequential and Random

Kumamoto University

Single and multiple pivots in two low-resource languages

- Result on Baseline Ja-Id
  - Baseline random obtained higher BLEU score, i.e., 12.17, compare to Baseline sequential, i.e., 11.96
  - Baseline random obtained higher perplexity score compare to Baseline sequential

Table 3.5: Ja-Id BLEU score on sequential data type

Systems	Cascade	Triangulation	LI	FI
Direct translation system				
Baseline		11.96		
Single pivot system				
JaId (English)	10.89	9.71	11.97	12.07
JaId (Myanmar)	9.37	8.71	11.91	12.27
JaId (Malay)	12.01	8.37	11.71	12.09
JaId (Filipino)	9.95	9.41	12.23	12.19

Table 3.6: Ja-Id BLEU score on random data type

Systems	Cascade	Triangulation	LI	FI
Direct translation system				
Baseline		12.17		
Single pivot system				
JaId (English)	10.81	9.10	12.18	12.22
JaId (Myanmar)	9.60	8.60	11.91	12.29
JaId (Malay)	11.81	9.25	12.22	12.05
JaId (Filipino)	9.68	9.62	12.09	11.99

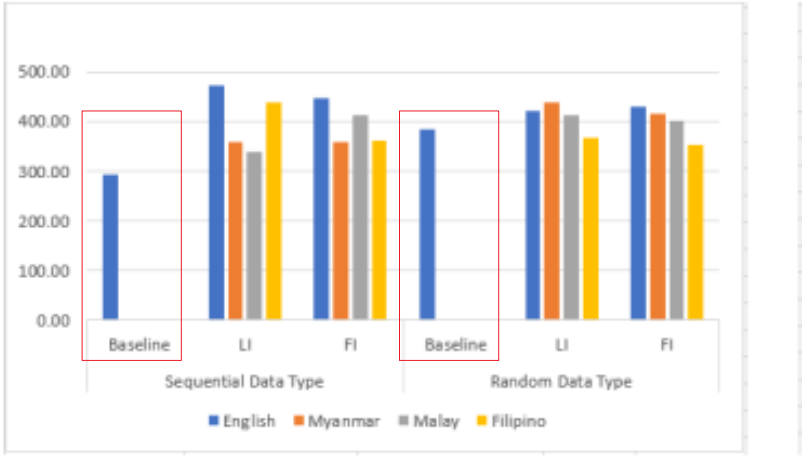


Figure 3.1: Perplexity Score of Ja-Id single pivot for LI and FI approaches.

Kumamoto University

Single and multiple pivots in two low-resource languages

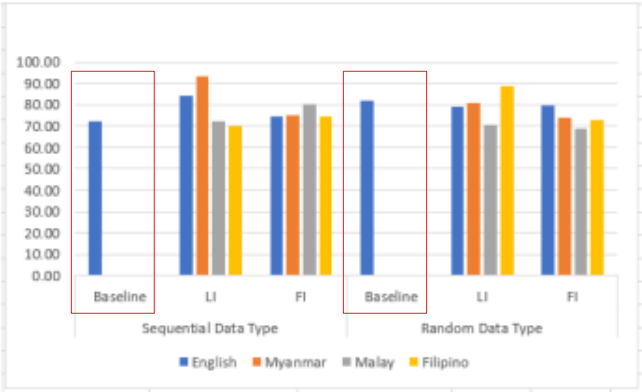
- Result on Baseline Id-Ja
  - Baseline random obtained higher BLEU score, i.e., 12.00, compare to Baseline sequential, i.e., 11.00
  - Baseline random obtained higher perplexity score compare to Baseline sequential

Table 3.7: Id-Ja BLEU score on sequence data type

Systems	Cascade	Triangulation	LI	FI
Direct translation system				
Baseline		11.00		
Single pivot system				
JaId (English)	12.07	8.26	12.65	12.05
JaId (Myanmar)	9.97	6.76	10.89	12.4
JaId (Malay)	12.18	6.76	12.2	11.87
JaId (Filipino)	10.36	7.28	12.06	12.2

Table 3.8: Id-Ja BLEU score on random data type

Systems	Cascade	Triangulation	LI	FI
Direct translation system				
Baseline		12.00		
Single pivot system				
JaId (English)	7.58	7.96	12.10	11.99
JaId (Myanmar)	10.32	6.51	12.84	12.88
JaId (Malay)	11.13	9.17	12.52	11.82
JaId (Filipino)	10.46	7.97	12.25	12.68



IFigure 3.2: Perplexity Score of Id-Ja single pivot for LI and FI approaches.

Kumamoto University

Single and multiple pivots in two low-resource languages

Results for single pivot

	Single pivot Ja-Id	Single pivot Id-Ja
Worse approach	Triangulation	Triangulation
Approach that improve the BLEU score	LI & FI	LI & FI
Pivot language choice	Seq: Myanmar Rnd: Malaysia & Myanmar	Seq: Myanmar Rnd: Myanmar

Table 3.5: Ja-Id BLEU score on sequential data type

Systems	Cascade	Triangulation	LI	FI
Direct translation system				
Baseline		11.96		
Single pivot system				
JaId (English)	10.89	9.71	11.97	12.07
JaId (Myanmar)	9.37	8.71	11.91	12.27
JaId (Malay)	12.01	8.37	11.71	12.09
JaId (Filipino)	9.95	9.41	12.23	12.19

Table 3.6: Ja-Id BLEU score on random data type

Systems	Cascade	Triangulation	LI	FI
Direct translation system				
Baseline		12.17		
Single pivot system				
JaId (English)	10.81	9.10	12.18	12.22
JaId (Myanmar)	9.60	8.60	11.91	12.29
JaId (Malay)	11.81	9.25	12.22	12.05
JaId (Filipino)	9.68	9.62	12.09	11.99

Table 3.7: Id-Ja BLEU score on sequence data type

Systems	Cascade	Triangulation	LI	FI
Direct translation system				
Baseline		11.00		
Single pivot system				
JaId (English)	12.07	8.26	12.65	12.05
JaId (Myanmar)	9.97	6.76	10.89	12.4
JaId (Malay)	12.18	6.76	12.2	11.87
JaId (Filipino)	10.36	7.28	12.06	12.2

Table 3.8: Id-Ja BLEU score on random data type

Systems	Cascade	Triangulation	LI	FI
Direct translation system				
Baseline		12.00		
Single pivot system				
JaId (English)	7.58	7.96	12.10	11.99
JaId (Myanmar)	10.32	6.51	12.84	12.88
JaId (Malay)	11.13	9.17	12.52	11.82
JaId (Filipino)	10.46	7.97	12.25	12.68

Kumamoto University

Single and multiple pivots in two low-resource languages

- Multiple pivots result
  - We use phrase tables of LI and FI from single pivot results
  - We combine four phrase tables
    - Without src-trg phrase table: All-LI, All-FI
    - With src-trg phrase table: Base-LI, Base-FI
  - The combination of All-LI and All-FI arranged by ascending and descending orders
    - Example: LI results from single pivot, i.e., 11.34 for EnPT, 12.21 for MyPT, 12.11 for MsPT, and 12.15 for FiPT.
    - All-LI ascending order: EnPT, MsPT, FiPT, MyPT,
    - All-LI descending order: MyPT, FiPT, MsPT, and EnPT

Single and multiple pivots in two low-resource languages

Results for multiple pivots

	Multiple pivots of Ja-Id	Multiple pivots of Id-Ja
Approaches that improve the BLEU score	All-LI, All-FI, Base-LI	All-LI, All-FI, Base-LI, Base-FI
Phrase table orders	Descending	Ascending
Data type	Base-LI Random: +0.23 point	All-FI Sequence: +1.84 point
Perplexity score	Base-LI Random could reduce perplexity score	All-LI, All-FI, Base-LI, Base-FI could reduce perplexity score

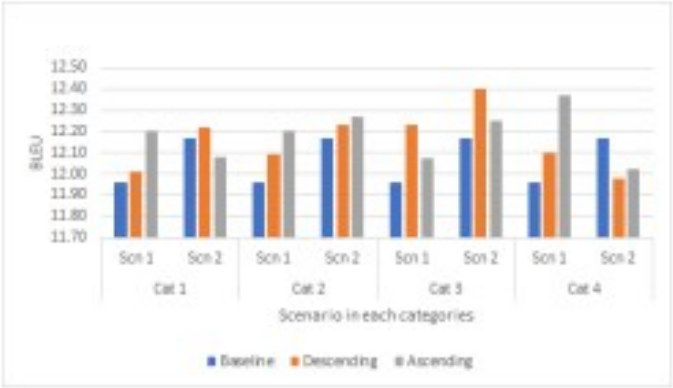


Figure 3.5: BLEU score for Ja-Id in multiple pivots.

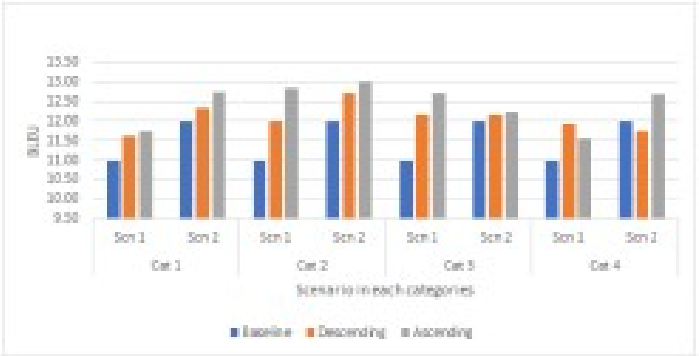


Figure 3.6: BLEU score for Id-Ja in multiple pivots.

Kumamoto University

Single and multiple pivots in two low-resource languages

Table 3.9: Best BLEU score in Baseline, single and multiple pivots for Ja-Id

Scenario No	Baseline	Single Pivot				Multiple Pivots	
		Cascade	Triangulate	Interpolate	Fillup Interpolation	Desc	Asc
Scenario 1	11.96	12.01 (MS)	9.71 (EN)	12.21 (MY)	12.27 (MY)	12.23 (Cat 3)	12.37 (Cat 4)
Scenario 2	12.17	11.81 (MS)	9.62 (FI)	12.22 (MS)	12.29 (MY)	12.40 (Cat 3)	12.27 (Cat 2)

Table 3.10: Best BLEU score in baseline, single and multiple pivots for Indonesia to Japanese

Scenario No	Baseline	Single Pivot				Multiple Pivots	
		Cascade	Triangulate	Interpolate	Fillup Interpolation	Desc	Asc
Scenario 1	11.00	12.18 (MS)	8.26 (EN)	12.03 (MY)	12.40 (MY)	12.15 (Cat 3)	12.84 (Cat 2)
Scenario 2	12.00	11.13 (MS)	9.17 (MS)	12.84 (MY)	12.88 (MY)	12.74 (Cat 2)	13.02 (Cat 2)

Proposed approach

1. Word reordering in multiple pivots for Japanese to Indonesian (Ja-Id)

2. The comparison of phrase table combination for Kazakh to English (Kk-En) and Japanese to Indonesian (Ja-Id)

Word reordering in multiple pivots for Ja-Id

- Word reordering in multiple pivots for Japanese to Indonesian (Ja-Id)
  - The SMT model is known as it does not work for language pairs that have different word order. (Bisazza and Federico, 2016; Isozaki et al., 2012; Simon and Purwarianti, 2013).
  - We also find this issue in our previous experiment, i.e., the generated text of Indonesian from our multiple pivots experiment followed the Japanese sentence structure.
  - Approaches for word order issue, i.e., pre-ordering, post-ordering, and word ordering as part of decoding process

s= Jishin[N] wa[PRT] '[PCT] nantō[N] Ajia[N] wo[PRT] kaimetsu[N] sa[V] se[AUXV] ta[AUXV] 2004[N] nen[N] no[PRT] indo[N] yō[SUF] dai[PRE]
jishin[N] ga[PRT] oso[V] tsu[TAIL] ta[AUXV] hi[N] kara[PRT] chōdo[ADV] ni[N] nen[N] go[SUF] ni[PRT] oki[V] ta[AUXV] '[PCT]

t= bahwa[SC] gempa[NN] Tenggara[NNP] '[Z] yang[SC] telah[MD] menghancurkan[VB] Asia[NNP] tahun[NN] 2004[CD] Samudera[NNP]
Hindia[NNP] gempa[NN] melanda[VB] besar[JJ] dari[IN] tanggal[NN][VB] yang[SC] hanya[RB] 2[CD] tahun[NN] setelah[SC] terjadi[VB] '[Z]
(that the Southeastern earthquake, which destroyed Asia in the 2004 Indian Ocean earthquake struck a large date from only 2 years after it occurred)

Kumamoto University

Experimental setup

- Without reordering (WoR)
  - → Taken from single and multiple pivots Ja-Id experimental result
  - → Taken from Triangulation, LI, multiple pivots of ascending
- With reordering (WR)
  - → Pre-ordering Japanese dataset by Lader (Neubig et al., 2012)
  - → Systems: direct translation as baseline, one pivot system (single pivot), two pivot systems, three pivots systems, four pivot systems.

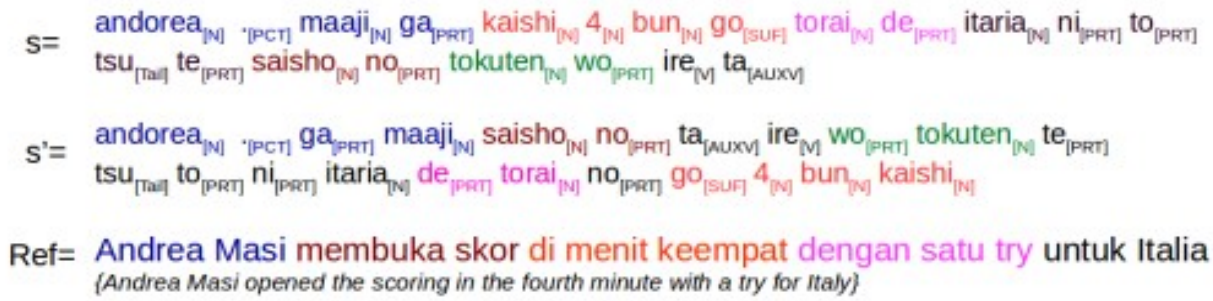


Figure 4.1: Word ordering of Ja sentence structure, i.e., SOV, into Indonesian sentence structure, i.e., SVO, using Lader.

Word reordering in multiple pivots for Ja-Id

- Proposed approach in src-pvt phrase table
  - We build the extended phrase table of src-pvt
  - The extended phrase table is a merging phrase table from two symmetrization techniques, viz., *gdfand* and *tgttosrc*
  - Based on the result that the unknown words of *gdfand* have candidate phrase pair in the tgttosrc
- Proposed approach step
  - We construct two pivot systems, viz., src-pvt *gdfand* and src-pvt *tgttosrc*
  - We sorted the unknown word of generated text from src-pvt *gdfand*
  - We query the unknown word from src-pvt *tgttosrc*
  - We merged the src-pvt *gdfand* and src-pvt *filtered* phrase table

Table 4.3: BLEU scores of src-pvt extended phrase table in WR experiment

Language pair	BLEU scores	
	gdfand phrase table	extended phrase table
Ja-En	8.11	8.14
Ja-Ms	7.60	7.56
Ja-Fi	7.95	8.01
Ja-My	4.50	4.45

Kumamoto University

Word reordering in multiple pivots for Ja-Id

WoR experimental result

- The Triangulation obtained the lowest BLEU score
- The multiple pivots of JaId(EnMsFiMy) outperformed the Baseline and LI, except JaId-My
- The multiple pivots of Baseline+(EnMsFiMy) outperformed the Baseline and Triangulation, however poorly compare to LI
- We analyze the decline of Baseline+(EnMsFiMy) BLEU score compare to JaId(EnMsFiMy)
  - ➔ Phrase table evaluation
    - Switch the phrase table: 1,041,599
  - ➔ Feature functions
    - Baseline+(EnMsFiMy) obtained lower feature functions, i.e., distortion weight, language model weight, translation model weight.

Kumamoto University

Word reordering in multiple pivots for Ja-Id

- WoR experimental result

Table 4.1: BLEU scores of single and multiple pivots in WoR experiments

Single pivot			Multiple pivot	
JaId			11.96	
Language Pair	Triangulation	LI	Language Pair	LI
JaId-En	9.71	11.34	JaId (EnMsFiMy)	12.20
JaId-My	8.71	12.21	Baseline + (EnMsFiMy)	12.07
JaId-Ms	8.37	12.11		
JaId-Fi	9.41	12.15		

Table 4.2: Generated text examples of single and multiple pivots in WoR experiments

Source (Ja)	地震は、南東アジアを壊滅させた2004年のインド洋大地震が襲った日からちょうど二年後に起きた。		
Language pair	Approach	BLEU score	Translation output
JaId (En)	Single pivot - Triangulation	9.71	gempa アジア tenggara, dan mereka untuk 壊滅 Samudra Hindia tahun 2004, menghantam gempa besar dari hanya dua tahun setelah 起き.
JaId (My)	Single pivot - LI	12.21	bahwa gempa Tenggara, yang telah menghancurkan Asia tahun 2004 Samudera Hindia gempa melanda besar dari tanggal yang hanya 2 tahun setelah terjadi.
JaId (EnMsFiMy)	Multiple pivots - LI	12.20	bahwa gempa Tenggara, yang telah menghancurkan Asia tahun 2004 Samudera Hindia gempa melanda besar dari tanggal yang hanya 2 tahun setelah terjadi.
Baseline + (EnMsFiMy)	Multiple pivots - LI	12.07	gempa SNT.57162.18909 tenggara, yang telah menghancurkan Asia tahun 2004 gempa melanda besar Samudera Hindia, yang hanya dari hari kedkecualiua terjadi pada tahun.

Kumamoto University

Word reordering in multiple pivots for Ja-Id

- WR experimental result
  - The Triangulation approach obtained the lowest BLEU score, however the generated text of one pivot WR experiment significantly change by means that it become more understand compared to one pivot WoR experiment.
  - The result shows that by combining more numbers of pivot languages, then the BLEU score gradually improved.
  - The generated text become more understandable, however the generated text of each system has the same result (text)

Word reordering in multiple pivots for Ja-Id

- WoR experimental result

Table 4.4: BLEU scores of single and multiple pivots in WR experiment

One pivot language		Two pivot language		Three pivot language		Four pivot language	
Language pair	Triangulation	Language Pair	LI	Language pair	LI	Language pair	LI
JaId				6.75			
JaId-En	5.99	JaId (EnMs)	6.92	JaId (EnMsFi)	6.94	JaId (MsEnFiMy)	7.15
JaId-Ms	6.30	JaId (EnFi)	6.29	JaId (EnMsMy)	6.98		
JaId-Fi	5.05	JaId (EnMy)	6.49	JaId (EnFiMy)	6.59		
JaId-My	3.16	JaId (MsFi)	6.73	JaId (MsFiMy)	6.85		
		JaId (MsMy)	6.46				

Table 4.5: Generated text examples of single and multiple pivots in WR experiment.

Source (Ja)	地震は、ちょうど 2 年後から日たった震インド洋大地震がの年 2004 たせさ壊滅南東アジアを。		
Language pair	Approaches	BLEU score	Translation Output
JaId-Ms	One pivot –Triangulation	6.30	gempa bumi tersebut terjadi hanya 2 tahun setelah dari hari melanda India besar gempa bumi pada tahun 2004 telah menghancurkan Tenggara Asia.
JaId (EnMs)	Two pivot –LI	6.92	gempa terjadi hanya 2 tahun setelah dari hari melanda India gempa besar pada tahun 2004 yang telah menghancurkan Asia selatan.
JaId (EnMsMy)	Three pivot –LI	6.98	gempa terjadi hanya 2 tahun setelah dari hari melanda India gempa besar pada tahun 2004 yang telah menghancurkan Asia Selatan.
JaId (MsEnFiMy)	Four pivot –LI	7.15	gempa terjadi hanya 2 tahun setelah dari hari melanda India gempa besar pada tahun 2004 yang telah menghancurkan Asia Selatan.

The comparison of phrase table combination for Kk-En and Ja-Id

Background

- The translation quality is influenced by Translation Model (Tian et al., 2014)
- The phrase table contains of phrase pairs that is extracted from word alignment by using symmetrization techniques such as grow-diag-final-and (gdfand), grow, final-grow, grow-diag, intersection, union, srctotgt, and tgttosrc.
- The pivot approach consist of three language pairs, viz., src-trg, src-pvt and pvt-trg, that uses *gdfand* as standard symmetrization.
- The various symmetrization have obtained different BLEU score on language pair in high resource languages (Koehn et al., 2005, Thoudam, 2015, Sara Styme et al., 2014).
- It needs to explore the symmetrization to know which symmetrization obtained high BLEU score in src-trg, src-pvt, and pvt-trg

The comparison of phrase table combination for Kk-En and Ja-Id

- Experimental setup
  - Language pairs
    - Kk→ En and Ja → Id
    - Kk → En taken from preliminary work, i.e., single pivot in Kk-En
    - Ja → Id taken from preliminary work, i.e., single and multiple pivots in Ja-Id
  - Language Model orders
    - 3-gram
    - 5-gram
  - Direct System Experiment (DSE)
    - Direct translation in src-trg, src-pvt, and pvt-trg
    - We explore five symmetrization, i.e., gdfand, intersection, union, srctotgt, tgttosrc

The comparison of phrase table combination for Kk-En and Ja-Id

- Interpolation System Experiment (ISE)
  - → Std-ISE (Standard Interpolation System Experiment)
    - We use gdfand symmetrization
  - → F-ISE (First-best Interpolation System Experiment)
    - We use first-high symmetrization obtained from DSE
  - → S-ISE (Second-best Interpolation System Experiment)
    - We use second-best symmetrization obtained from DSE

The comparison of phrase table combination for Kk-En and Ja-Id

- Experimental result in DSE
  - The LM05 systems obtained higher BLEU score compare to LM03
  - Several results show that the non-standard symmetrization obtained a higher BLEU score compare to the standard one.
  - We investigate the different BLEU scores
    - ➔ Phrase translation parameters obtained different BLEU scores, despite we use the same word alignment tools, i.e., MGIZA++.
    - ➔ The phrase translation parameters is scores obtained from scoring functions
      - inverse phrase translation probability (p(t|s)),
      - inverse lexical weighting (lex(t|s))
      - direct phrase translation probability (p(s|t))
      - direct lexical weight (lex(s|t)).

The comparison of phrase table combination for Kk-En and Ja-Id

Table 4.8: The obtained BLEU scores of Direct System Experiments (DSE). Results in bold indicate the first highest translation quality, while those in italic indicate the second highest translation quality.

Language pair - System	BLEU scores				
	gdfand	intersection	union	srctotgt	tgttosrc
Kk-En LM03	*3.08*	2.05	3.07	2.51	**3.36**
Kk-En LM05	*3.42*	2.26	3.28	2.77	**3.56**
Kk-Ru LM03	**6.22**	4.98	4.31	*5.41*	5.10
Kk-Ru LM05	**6.49**	5.17	4.35	*5.64*	5.56
Ru-En LM03	**4.77**	0	2.92	*4.09*	3.12
Ru-En LM05	**4.63**	0	2.73	*3.80*	2.85
Ja-Id LM03	**11.96**	10.54	9.55	9.79	*11.63*
Ja-Id LM05	**12.2**	10.47	9.43	9.82	*12.04*
Ja-Ms LM03	**12.95**	10.09	10.23	10.46	*12.65*
Ja-Ms LM05	**13.24**	11.06	10.17	10.54	*12.93*
Ms-Id LM03	**35.07**	34.66	34.90	34.52	*34.99*
Ms-Id LM05	*35.04*	34.75	34.89	34.62	**35.14**

Table 4.10: Example of phrase translation parameter scores in Kk-En LM05. Results in bold indicates the score is higher.

Phrase-pair	Phrase translation parameters	Symmetrization	
		gdfand	tgttosrc
2007 жылдан бастап ||| since 2007,	Inverse phrase translation probability (p(f|e))	0.5	0.5
	Inverse lexical weighting (lex(f|e))	0.000930714	**4.8791e-05**
	Direct phrase translation probability (p(e|f))	0.5	0.333333
	Direct lexical weighting (lex(e|f))	0.00596183	**0.0128321**

Kumamoto University

The comparison of phrase table combination for Kk-En and Ja-Id

Table 4.9: The symmetrization technique candidate for ISE. Results in (1) is a symmetrization technique for F-ISE, and (2) is for S-ISE, when doing phrase table combination

Kk-En				Ja-Id			
LM order	Lang pair	(1)	(2)	LM order	Lang pair	(1)	(2)
LM03	Kk-En	tgttosrc	gdfand	LM03	Ja-Id	gdfand	tgttosrc
	Kk-Ru	gdfand	srctotgt		Ja-Ms	gdfand	tgttosrc
	Ru-En	gdfand	srctotgt		Ms-Id	gdfand	tgttosrc
LM05	Kk-En	tgttosrc	gdfand	LM05	Ja-Id	gdfand	tgttosrc
	Kk-Ru	gdfand	srctotgt		Ja-Ms	gdfand	tgttosrc
	Ru-En	gdfand	srctotgt		Ms-Id	tgttosrc	gdfand

The comparison of phrase table combination for Kk-En and Ja-Id

- Experimental result in ISE
  - The LM05 systems obtained higher BLEU score compare to LM03
  - The F-ISE is a competitive approach because it can improve the BLEU score in Kk → En
    - → The translation parameters of phrase pair F-ISE obtained higher compare to others
  - The Baseline is outperformed the F-ISE in Ja → Id
    - → The translation parameters of phrase pair Baseline obtained higher compare to F-ISE
  - The phrase table size does not directly affect the improvement of the BLEU score.

The comparison of phrase table combination for Kk-En and Ja-Id

- Language Model orders
  - The longer LM order of Kk-En, i.e., LM05, obtained lower perplexity score in all systems.
  - The longer LM order could not obtain a lower perplexity score in Ja-Id.
    - Target monolingual size of Kk→En is thirteen times bigger, i.e., 114,375, than Ja→ Id, i.e., 8,500
    - We use the same dataset for training and LM in Ja-Id, i.e., *DataALT.01.jp-id.SP.true.id: 8,500*
    - The Kk-En use different datasets for training, i.e., i.e., *news-commentary-v14.en-kk.en*:*9,600* and LM, i.e., *news-commentary-v14.en-ru.en: 114,375.*
- Generated text
  - Wrong word position because it is followed the source sentence pattern

The comparison of phrase table combination for Kk-En and Ja-Id

Table 4.11: BLEU scores of the system

Language Model	Direct translation	Std-ISE	F-ISE	S-ISE
Kk-En				
LM03	3.08	3.08	3.43	3.09
LM05	3.42	3.42	3.64	3.42
Ja-Id				
LM03	11.96	12.07	12.07	11.16
LM05	12.20	12.08	12.08	11.26

Table 4.13: Phrase translation scores of Kk-En LM05. Results in bold indicates the score is higher.

Phrase-pair	Phrase translation parameters	Phrase translation scores		
		Baseline	Std-ISE	F-ISE
алу экономикалық ||| maintain economic	Inverse phrase translation probability (p(f|e))	0.25	0.2484	0.0014
	Inverse lexical weighting (lex(f|e))	0.0011	0.0011	**0.0013**
	Direct phrase translation probability (p(e|f))	0.5	0.4968	**0.9723**
	Direct lexical weighting (lex(e|f))	0.0002	0.0002	**0.7641**
жаңа бір маңызды саясатты ||| important new policies	Inverse phrase translation probability (p(f|e))	0.3333	0.3312	0.3241
	Inverse lexical weighting (lex(f|e))	0.0001	0.0001	0.0006
	Direct phrase translation probability (p(e|f))	0.25	0.2484	**0.972348**
	Direct lexical weighting (lex(e|f))	0.1046	0.1039	**0.1091**

Table 4.12: Phrase translation scores of Ja-Id LM05. Results in bold indicates the score is higher.

Phrase-pair	Phrase translation parameters	Phrase translation scores		
		Baseline	Std-ISE	F-ISE
から 強制 ||| rumahnya digerebek	Inverse phrase translation probability (p(f|e))	**0.3333**	0.2946	0.2972
	Inverse lexical weighting (lex(f|e))	**0.0002**	0.0001	0.0001
	Direct phrase translation probability (p(e|f))	**0.3333**	0.2946	0.2949
	Direct lexical weighting (lex(e|f))	**1.23E-07**	1.09E-07	1.09E-07
から 情報 を 受け取る ||| menerima informasi dari	Inverse phrase translation probability (p(f|e))	**0.25**	0.2209	0.2229
	Inverse lexical weighting (lex(f|e))	**0.0024**	0.0018	0.0018
	Direct phrase translation probability (p(e|f))	**1**	0.8839	0.8847
	Direct lexical weighting (lex(e|f))	**0.1019**	0.0899	0.0901

The comparison of phrase table combination for Kk-En and Ja-Id

Table 4.14: Phrase table size of the system

Language Model	Baseline	Standard approach (gdfand)	Initiative approach	
			F-ISE	S-ISE
Kk-En				
LM03	723,960	742,948	323,850	730,544
LM05	723,960	742,948	323,850	730,544
Ja-Id				
LM03	875,038	935,717	935,717	750,449
LM05	875,038	935,717	925,732	764,219

Table 4.15: Perplexity scores of the system

Language Model	Baseline	Standard approach (gdfand)	Initiative approach	
			F-ISE	S-ISE
Kk-En				
LM03	148.21	148.18	284.05	148.39
LM05	93.41	115.90	206.15	115.86
Ja-Id				
LM03	309.32	310.25	310.25	310.09
LM05	403.13	411.48	414.46	386.94

s= Sono[Adn] 2004[noun] nen[noun] no[particle] jishin[noun] wa[particle] magunichūdo[noun] 9.1[noun] to[particle] kiroku[noun] sa[verb] re[verb] "[sym] 33[noun] fito[noun] no[particle] taka[adj] sa[noun] ni[particle] tassuru[verb] nami[verb] o[particle] tomonau[verb] tsunami[noun] o[particle] hikiokoshi[verb] ta[aux verb] "[sym]

t= tahun[NN] 2004[CD] tersebut[PR,VB] bahwa[SC] gempa bumi[NN] dengan[SC,IN] kekuatan[NN] 9.1[CD] dan[CC] 33[CD] kaki[NN] "[Z] yang[SC] mencapai[VB] ketinggian[NN] menyebabkan[VB] tsunami[X] yang[SC] gelombang[NN] "[Z]

(in 2004 that an earthquake with a magnitude of 9.1 and 33 feet, which reached a height caused a tsunami of waves.)

Figure 4.3: Sentence structure for Ja-Id, taken from LM05 F-ISE

Kumamoto University

Conclusion and future work

- Conclusion
  - The interpolation system of single pivot outperformed the direct translation. The non-English as pivot languages could be considered as pivot language because it could improve the translation quality.
  - The more number of pivot languages (multiple pivots) could improve the BLEU score compare to the direct translation.
  - The phrase table order, i.e., ascending and descending, could be considered when multiple pivots employed.
  - The data type, i.e., sequence and random, could be considered to improve the BLEU score
  - The longer LM order, i.e., 5-gram, obtained higher BLEU score compare to the short one, i.e., 3-gram

Conclusion and future work

- The pre-ordering should be employ for language pair that have different word order
- The improvement of src-pvt and pvt-trg should be considered to improve the pivot BLEU score
  - ➔ The extending src-pvt phrase table
  - ➔ The exploration of non-standard symmetrization
    - Find the first-best (F-ISE)
    - Find the second-best (S-ISE)

Future work

- The extending pvt-trg phrase table
- Increasing Indonesian target monolingual target size
- Implement the F-ISE in multiple pivots

Thank you