Centre for Next Generation Localisation

Low-Resource Machine Translation Using MaTrEx: The DCU Machine Translation System for IWSLT 2009

Yanjun Ma, Tsuyoshi Okita, Özlem Çetinoğlu, Jinhua Du, Andy Way, Dublin City University, CNGL/School of Computing

National Development Plan 2007 - 2013

science foundation ireland

Dublin City University University College Dublin University of Limerick Trinity College Dublin

Table Of Contents

IWSLT Rationale

- IWSLT pursues research aspects: No additional resources other than corpora provided.

> *... certain gains in performance were triggered by better suited language resources (engineering aspects) or by improvements in the underlying decoding algorithms and statistical models (research aspects). (IWSLT organizer)*

MaTrEx: Low-Resource Machine Translation

- MaTrEx for Low-Resource MT
  - Word Lattice
    - Rational: We have space to investigate various segmentation in Chinese and Turkish.
  - Noise Reduction
    - Rational: There would be various paraphrases, multiword expressions, non-literal translations included in bitext.
  - Multiple System Combination
  - Case and Punctuation Restoration

MaTrEx: Low-Resource Machine Translation

- MaTrEx for Low-Resource MT
  - Word Lattice
    - Rational: We have space to investigate various segmentation in Chinese and Turkish.
  - Noise Reduction
    - Rational: There would be various paraphrases, multiword expressions, non-literal translations included in bitext.
  - Multiple System Combination
  - Case and Punctuation Restoration
- MaTrEx participated in 2006/7/8/9, Turkish first time

IWSLT 2009 Corpora

- BTEC task (Basic Travel Expression Corpus) and CHALLENGE task (which uses Spoken Language Databases corpus).
  - BTEC task: Chinese-English and Turkish-English
  - CHALLENGE task: Chinese-English and English-Chinese

IWSLT 2009 Corpora

- BTEC task (Basic Travel Expression Corpus) and CHALLENGE task (which uses Spoken Language Databases corpus).
  - BTEC task: Chinese-English and Turkish-English
  - CHALLENGE task: Chinese-English and English-Chinese

	train set	dev set	test set
BT-TR-EN	27,972	506 ($\times$16)	469
BT-ZH-EN	47,098	507 ($\times$16)	469
CH-ZH-EN	75,231	489 ($\times$7)	405
CH-EN-ZH	39,228	210 ($\times$4)	393

Table: Parallel corpus size of IWSLT 2009 (Only our participated tasks)

Table Of Contents

Word Lattice

- Speech recognition: first determine the best word segmentation and perform decoding (the accoustic signal underdetermines the choice of source word sequence).

$$\hat{v}_1^K = \arg\max_{v_1^K, K}\{P(v_1^K | f_1^I)\}, \qquad \hat{e}_1^J = \arg\max_{e_1^J, J}\{P(e_1^J | \hat{v}_1^K)\}$$

Word Lattice

- Speech recognition: first determine the best word segmentation and perform decoding (the accoustic signal underdetermines the choice of source word sequence).

$$\hat{v}_1^K = \arg\max_{v_1^K, K}\{P(v_1^K|f_1^I)\}, \qquad \hat{e}_1^J = \arg\max_{e_1^J, J}\{P(e_1^J|\hat{v}_1^K)\}$$

- Word lattice-based approach in SMT: to allow the MT decoder to consider all possibilities for f by encoding the alternatives compactly as a word lattice. [Xu et al., 2005][Bertoldi et al., 2007][Dyer et al., 2008][Ma and Way, EACL2009].

$$\begin{aligned}\hat{e}_1^J &= \arg\max_{e_1^J, J}\{\max_{v_1^K, K} P(e_1^J, v_1^K|f_1^I)\} = \arg\max_{e_1^J, J}\{\max_{v_1^K, K} P(e_1^J)P(v_1^K|e_1^J, f_1^I)\} \\ &\simeq \arg\max_{e_1^J, J}\{\max_{v_1^K, K} p(e_1^J)p(v_1^K|f_1^I)p(v_1^K|e_1^J)\}\end{aligned}$$

Word Lattice: Generation (Chinese)

Chinese (word boundaries are not orthographically marked)
在门厅下面。我这就给您拿一些。

(zai men ting xia mian. wo zhe jiu gei nin na yi xie)

Word Lattice: Generation (Chinese)

Chinese (word boundaries are not orthographically marked)
在门厅下面。我这就给您拿一些。

(zai men ting xia mian. wo zhe jiu gei nin na yi xie)

1. Manual segmentation
   在_门厅_下面_。_我_这_就_给_您_拿_一些_。

Word Lattice: Generation (Chinese)

Chinese (word boundaries are not orthographically marked)
在门厅下面。我这就给您拿一些。

(zai men ting xia mian. wo zhe jiu gei nin na yi xie)

1. Manual segmentation
   在_门厅_下面_。_我_这_就_给_您_拿_一些_。
2. LDC segmentation
   在_门_厅_下面_。_我_这_就_给_您_拿_一些_。

Word Lattice: Generation (Chinese)

Chinese (word boundaries are not orthographically marked)
在门厅下面。我这就给您拿一些。

(zai men ting xia mian. wo zhe jiu gei nin na yi xie)

1. Manual segmentation
   在_门厅_下面_。_我_这_就_给_您_拿_一些_。
2. LDC segmentation
   在_门_厅_下面_。_我_这_就_给_您_拿_一些_。
3. Character-based segmentation
   在_门_厅_下_面_。_我_这_就_给_您_拿_一_些_。

Word Lattice: Generation (Turkish)

Turkish (rich morphology language)
Bu mevsimin en yeni rengi ne?

1. lowercased original data
    - each word is a segment
      bu mevsimin en yeni rengi ne ?

Word Lattice: Generation (Turkish)

Turkish (rich morphology language)
Bu mevsimin en yeni rengi ne?

1. lowercased original data
   - each word is a segment
     bu mevsimin en yeni rengi ne ?
2. morphologically analyzed [Oflazer, 94] and disambiguated [Sak, 07], and reduced analysis, i.e., only informative morphemes are kept [Oflazer].
   - each analysis is a segment
     bu+Det mevsim+Noun+Gen en+Adverb yeni+Adj renk+Noun+P3sg ne+Adverb ?

Word Lattice: Generation (Turkish)

Turkish (rich morphology language)
Bu mevsimin en yeni rengi ne?

1. lowercased original data
   - each word is a segment
     bu mevsimin en yeni rengi ne ?
2. morphologically analyzed [Oflazer, 94] and disambiguated [Sak, 07], and reduced analysis, i.e., only informative morphemes are kept [Oflazer].
   - each analysis is a segment
     bu+Det mevsim+Noun+Gen en+Adverb yeni+Adj renk+Noun+P3sg ne+Adverb ?
   - each morpheme is a segment
     bu Det mevsim Noun Gen en Adverb yeni Adj renk Noun P3sg ne Adverb ?

Word Lattice: An Example

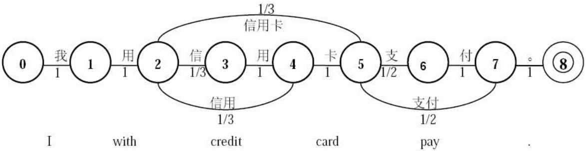

Figure: An example of a word lattice for a Chinese sentence

- Arc: segmented words.
- Numbers at arc: transition probabilities (1, 1/3, 1/2, and so forth).

Noise Reduction in MT

- Noise: statistical property
  - Noise reduction for phrase alignment [Tomeh et al., 2009]
- Outlier: dependent on underlying machine learning algorithm
  - Noise reduction for word alignment [Okita, ACL09SRW]
- Noise: defined by similarity measure (In sentence alignment, the removal of some particular sentence does not matter the quality in later stages)
  - Noise reduction for sentence alignment [Utiyama and Isahara, 2003]

Noise Reduction [Okita, ACL09SRW]

- (Training Phase) We let our MT systems learn by training set.

c' est la vie .	MT Systems	that is life .
je t' aime .	⇒ *Noisy Channel* ⇒	i love you .
elle est petite .		she is small .

Noise Reduction [Okita, ACL09SRW]

- (Training Phase) We let our MT systems learn by training set.

c' est la vie . je t' aime . elle est petite .	MT Systems $\Rightarrow$ *Noisy Channel* $\Rightarrow$	that is life . i love you . she is small .

- (Test Phase) We can expect if we translate our training set our MT systems learn most of them in good faith (considering a bit about generalisation error).

c' est la vie . je t' aime . elle est petite .	above trained $\Rightarrow$ MT systems $\Rightarrow$??	that is life . i love you . she is small .

Noise Reduction [Okita, ACL09SRW]

- (Training Phase) We let our MT systems learn by training set.

c' est la vie .	MT Systems	that is life .
je t' aime .	$\Rightarrow$ *Noisy Channel* $\Rightarrow$	i love you .
elle est petite .		she is small .

- (Test Phase) We can expect if we translate our training set our MT systems learn most of them in good faith (considering a bit about generalisation error).

c' est la vie .	above trained	that is life .
je t' aime .	$\Rightarrow$ MT systems $\Rightarrow$??	i love you .
elle est petite .		she is small .

- (Training Phase) We train our multiclass classifier by training set.

c' est la vie .	multiclass classifier	blue
je t' aime .	$\Rightarrow \Rightarrow$	red
elle est petite .		purple

Noise Reduction [Okita, ACL09SRW]

- (Training Phase) We let our MT systems learn by training set.

c' est la vie .	MT Systems	that is life .
je t' aime .	⇒ *Noisy Channel* ⇒	i love you .
elle est petite .		she is small .

- (Test Phase) We can expect if we translate our training set our MT systems learn most of them in good faith (considering a bit about generalisation error).

c' est la vie .	above trained	that is life .
je t' aime .	⇒ MT systems ⇒ ??	i love you .
elle est petite .		she is small .

- (Training Phase) We train our multiclass classifier by training set.

c' est la vie .	multiclass classifier	blue
je t' aime .	⇒ ⇒	red
elle est petite .		purple

- (Test Phase) We expect that multiclass classifier outputs similar color in our training set.

c' est la vie .	multiclass classifier	blue
je t' aime .	⇒ ⇒	red
elle est petite .		purple

Noise Reduction

- (Training Phase) 总共 是 多少 ? (zong gong shi duo shao) $\longrightarrow$ what does that come to ?

总共 是 多少 ? NULL ({ }) what ({ }) does ({ 1 2 3 }) that ({ }) come ({ }) to ({ }) ? ({ 4 })
what does that come to ? NULL ({ }) 总共 ({ 2 3 4 5 }) 是 ({ }) 多少 ({ 1 }) ? ({ 6 })

cause word alignment problem.

总共 是 多少 ? ||| what does that come to ? ||| · · · ||| · · · ||| 0.5 2.23258e-06 1 2.53525e-07 2.718 总共 是 多少 ||| what does that come to ||| · · · ||| · · · ||| 0.5 3.596e-06 1 2.62101e-07 2.718
总共 ||| total ||| (0) ||| (0) ||| 0.142857 0.0543478 0.125 0.0862069 2.718 是 ||| 's the ||| (0,1) ||| (0) (0) ||| 0.275862 0.0883644 0.00298954 0.000933415 2.718 多少 ||| what ||| (0) ||| (0) ||| 0.0480072 0.109269 0.254808 0.157088 2.718 ? ||| ? ||| (0) ||| (0) ||| 0.447633 0.620852 0.931172 0.967281 2.718

- (Test Phrase) 总共 是 多少 ? $\longrightarrow$ what 's the total ?

Noise Reduction

- Why is this noise reduction for word alignment?
  - 'word alignment + phrase extraction heuristics' is a compromise to solve a phrase alignment task [Marcu and Wong, 2002],
  - By definition, a word alignment task will not capture the *NtoM* mapping objects such as paraphrases, multi-word expressions and non-literal translations.
- (Heuristics in outlier detection literature): If we collect 'good points', we may be able to avoid outliers [Forsyth and Ponce, 2003].

Noise Reduction: Algorithm and Results

Algorithm 1 Good Points Algorithm

Step 1: Train word-based SMT, and translate all the sentences to get n-best lists.

Noise Reduction: Algorithm and Results

Algorithm 2 Good Points Algorithm

Step 1: Train word-based SMT, and translate all the sentences to get n-best lists.

Step 2: Obtain the sentence-based cumulative X-gram ($X \in \{1, \cdots, 4\}$) score $S_{WB,X}$.

Noise Reduction: Algorithm and Results

Algorithm 3 Good Points Algorithm

Step 1: Train word-based SMT, and translate all the sentences to get n-best lists.

Step 2: Obtain the sentence-based cumulative X-gram ($X \in \{1, \cdots, 4\}$) score $S_{WB,X}$.

Step 3: Train PB-SMT, and translate all training sentences to get n-best lists.

Noise Reduction: Algorithm and Results

Algorithm 4 Good Points Algorithm

Step 1: Train word-based SMT, and translate all the sentences to get n-best lists.
Step 2: Obtain the sentence-based cumulative X-gram ($X \in \{1, \cdots, 4\}$) score $S_{WB,X}$.
Step 3: Train PB-SMT, and translate all training sentences to get n-best lists.
Step 4: Obtain the sentence-based cumulative X-gram ($X \in \{1, \cdots, 4\}$) score $S_{PB,X}$.

Noise Reduction: Algorithm and Results

Algorithm 5 Good Points Algorithm

Step 1: Train word-based SMT, and translate all the sentences to get n-best lists.
Step 2: Obtain the sentence-based cumulative X-gram ($X \in \{1, \cdots, 4\}$) score $S_{WB,X}$.
Step 3: Train PB-SMT, and translate all training sentences to get n-best lists.
Step 4: Obtain the sentence-based cumulative X-gram ($X \in \{1, \cdots, 4\}$) score $S_{PB,X}$.
Step 5: Remove sentence pairs where $S_{WB,2} = 0$ and $S_{PB,2} = 0$.

Noise Reduction: Algorithm and Results

Algorithm 6 Good Points Algorithm

Step 1: Train word-based SMT, and translate all the sentences to get n-best lists.
Step 2: Obtain the sentence-based cumulative X-gram ($X \in \{1, \cdots, 4\}$) score $S_{WB,X}$.
Step 3: Train PB-SMT, and translate all training sentences to get n-best lists.
Step 4: Obtain the sentence-based cumulative X-gram ($X \in \{1, \cdots, 4\}$) score $S_{PB,X}$.
Step 5: Remove sentence pairs where $S_{WB,2} = 0$ and $S_{PB,2} = 0$.
Step 6: The remaining sentence pairs after removal in Step 5 are used to train the final PB-SMT systems.

Noise Reduction: Example of Detected Outliers

总共 是 多少 ?	what does that come to ?
服务台 的 号码 是 多少 ?	what number should i dial for information ?
它 在 星期几 开 ?	what days of the week does it take place ?
这 是 钥匙 。	the keys go here .
一 点 过 五 分 。	it 's five after one .

Table: Outliers for BTEC Chinese–English task by Good Point algorithm.

System Combination

- Minimum Bayes-Risk-Confusion Network (MBR-CN) framework [Kumar and Byrne, 2004][Du et al., WMT2008] (Work very well in our recent MT-eval campaigns).

$$\hat{e}_i = \arg\min_{e_i} \sum_{j=1}^{N} \{1 - BLEU(e_j, e_i)\}$$

- Confusion Network:
  - (backbone) output of MBR decoder, (other elements) other hypotheses are aligned by TER.
  - Features: 1) word posterior probability, 2) trigram and 4-gram target language model, 3) word length penalty, and 4) NULL word length penalty.
  - MERT is used to tune the weights of CN.

Case and Punctuation Restoration (1)

- Translation-based approach [Hassan et al., 07] (best system for Arabic-EN human evaluation)
  - Treating case / punctuation restoration as a translation task
    - source: lower-cased sentences
    - target: true-cased sentences (case restoration), text with punctuation (punctuation restoration)

Case and Punctuation Restoration (2)

- Punctuation restoration
  - Combination of translation-based approach and LM-based approach (by majority voting); If no solution can be found using this approach, we choose the first hypothesis proposed by the LM-based method).
- Case restoration
  - Translation-based approach.

Centre for Next Generation Localisation

Table Of Contents

Experimental Setup

- Baseline System: Standard log-linear PB-SMT system
  - word alignment by Giza++,
  - phrase extraction heuristics,
  - MERT (optimised by Bleu),
  - 5-gram language model with Kneser-Ney smoothing by SRILM, and
  - Moses [Koehn et al., 07] for decoding.
- System Combination
  - Joshua (Hierarchical Phrase-Based system) [Li et al., 09],
  - SAMT (Syntax-Based SMT nsystem) [Zollmann et al., 06].
- Additional Tools
  - LDC segmenter (Additional Chinese segmentation for word lattice),
  - Berkeley parser (required for Syntax-Based SMT systems),

Notation

GDF	grow-diag-final
INT	intersection
DS-GDF	noise reduction after grow-diag-final
Lattice	word lattice
HPB	hierarchical MT (joshua)
SAMT	syntax-based MT (SAMT)

BTEC Chinese–English translation

	PB-SMT			Lattice		HPB	SAMT	SCombo
	GDF	INT	DS	GDF	INT			
c/p	.3903	.3856	.3733	.4002	.3672	.3783	.3612	.4197
n c/p	.3808	.3717	.3617	.3811	.3463	.3614	.3466	.4135
OOV	139	90	191	40	6	139	141	48

Table: Performance of single systems and multiple system combination for BTEC Chinese–English translation (BLEU)

- sys combo 5 % increase than GDF.
- OOV

BTEC Turkish–English translation

	PB-SMT			Lattice		HPB	SAMT	SCombo
	GDF	INT	DS	GDF	INT			
c/p	.4831	.4656	.4591	.5233	.5247	.4711	.4708	.5593
n c/p	.4590	.4394	.4390	.5008	.5065	.4455	.4516	.5401
OOV	106	61	106	21	11	88	80	17

Table: Performance of single systems and multiple system combination for BTEC Turkish–English translation (BLEU)

- sys combo 7 % increase.

CHALLENGE Chinese–English translation

	PB-SMT			Lattice		HPB	SAMT	Combo
	GDF	INT	DS	GDF	INT			
crr c/p	.3169	.3278	.3143	.3436	.3335	.3148	.2978	.3689
n c/p	.3109	.3262	.3088	.3371	.3310	.3057	.2906	.3673
OOV	197	76	188	21	0	191	197	16
asr c/p	.2918	.2915	.2913	.2724	.2958	.2869	.2700	.3161
n c/p	.2789	.2825	.2752	.2660	.2861	.2744	.2536	.3064
OOV	158	96	153	5	5	157	154	5

Table: Performance of single systems and multiple system combination for CHALLENGE Chinese–English translation (BLEU)

CHALLENGE English–Chinese Results

	PB-SMT			HPB	SAMT	Combo
	GDF	INT	DS			
crr c/p	.3531	.3833	.3547	.3797	.3563	.3725
n c/p	.3555	.3885	.3570	.3832	.3613	.3757
OOV	99	32	91	102	101	38
asr c/p	.2970	.3264	.3138	.3332	.3088	.3273
n c/p	.2987	.3315	.3154	.3372	.3110	.3306
OOV	129	64	141	112	120	40

Table: Performance of single systems and multiple system combination for BTEC English–Chinese translation (BLEU)

- Sys combo decreases. This problem was investigated further [Du et al., ICASSP submitted].

Translation Example: Notation

1. PB
2. PB-INT
3. HIERO
4. SAMT
5. LATTICE
6. LATTICE-INT
7. DS-GDF
8. COMBO

Translation Example (1)

我 想 订 的 房间 住 两 天 是 十月 二十七日 和 二十八日 还 有 空房 吗 ?

1. I'd like to reserve a room for two nights, on October 二十七日 and 二十八日. Do you have any available rooms?
2. I'd like to reserve a room for two days, October 二十七日 and 二十八日 available?
3. I'd like to reserve a room for two nights, on October 二十七日 and 二十八日. Do you have any available rooms?
4. I'd like to reserve a room for two days, it's October 二十七日 and 二十八日. Do you have any available rooms?
5. I'd like to reserve a room for the twenty-seventh and twenty-eighth two Japanese yen available?
6. I'd like to reserve a room for two for the twenty-seventh and twenty-eighth of Japanese yen available?
7. I'd like to reserve a room for two days is on October 二十七日 and 二十八日. Do you have any rooms available?
8. I'd like to reserve a room for two for the twenty-seventh and do you have any available rooms?

Translation Example (2)

我 的 名字 是 铃木 直子 。

1. My name is Naoko Suzuki.
2. My name is Naoko Suzuki.
3. My name is Naoko Suzuki.
4. My name is Naoko Suzuki.
5. My name is Naoko Suzuki.
6. My name is Naoko Suzuki.
7. My name is Naoko Suzuki.
8. My name is Naoko Suzuki.

Centre for Next Generation Localisation

Translation Example (3)

嗯 我 还 一 个 问题 要 问 。

1. I'd also like to know a problem.
2. Well, I have a question.
3. Ask me, I'd like a question.
4. Yes, and I'd like to ask a problem.
5. I have a problem.
6. I'd like to ask you a question.
7. I'd also like to know a problem.
8. I'd like to have a problem.

TranslationExample (4)

我 用 万事达信用卡 付款 。

1. Can I use a Master card to pay.
2. I use a Master Card.
3. I use a Master card to pay.
4. I want to pay by MasterCard, please.
5. I use a Master Card.
6. I use a Master Card.
7. I want to pay by MasterCard, please.
8. I want to use a Master Card.

Translation Example (5)

是的 没错 。

1. Yes, 没错.
2. Yes, 没错.
3. Yes, 没错.
4. Yes, 没错.
5. Yes, that’s correct.
6. Yes, that’s right.
7. Yes 没错.
8. Yes, right.

Translation Example (6)

我 拿 北纬 太 贵 了 。

1. I have 北纬 too expensive.
2. I 北纬 too expensive.
3. I have 北纬 too expensive.
4. I have 北纬 too expensive.
5. I have north Montreal too expensive.
6. I Montreal north too expensive.
7. I'd like 北纬 too expensive.
8. I have north too expensive.

System Combination (Problem)

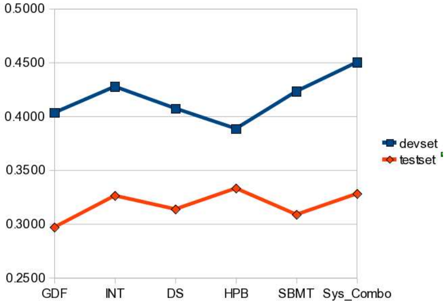

Figure: Performance of the systems on development set and test set

Why Noise Reduction Did Not Work? (1)

- (At first sight) Too much removal of sentences, 10-15 %, caused the problem (Our experiences for European language pairs, 3-5 % worked well).
- Close look at parallel corpus:
  - There are indeed a lot of duplicate pairs of sentences (this might cause the similar effect of noise reduction algorithm; removal vs duplication).

	train set	pure train set	redundancies
BT-TR-EN	27,972	26,970	3.0 %
BT-ZH-EN	47,098	43,657	12.2 %
CH-ZH-EN	75,231	69,680	4.0 %
CH-EN-ZH	39,228	38,227	12.0 %

Table: Redundancies in Parallel corpus

Why Noise Reduction Did Not Work? (2)

Sentence duplication algorithm [Okita, CLUKI09].

- motivated by statistics, make the tails of a probability distribution heavier.
- We tuned parameter by trial and error.

Algorithm 7 Sentence Duplication Algorithm

Step 1: Conditioned on a sentence length pair (l_e, l_f), we count the numbers of them. We calculate the ratio $r_{i,j}$ of this number over the number of all sentences.
Step 2: If this ratio $r_{i,j}$ is under the threshold X, we duplicate N times.

Why Noise Reduction Did Not Work? (3)

	train set	pure train set	noise reduction	removal
BT-TREN	.4831	.4478	.4611	7.1 %
BT-ZHEN	.3903	.3750	.3741	10.4 %
CH-ENZH	.3169	.2847	.3011	10.6 %
CH-ZHEN	.3531	.3154	–	9.5 %
	organizer	baseline	ours	

Table: BLEU score of original / non-redundant train set / noise reduced for non-redundant train set (PB-SMT by GDF setting).

- After applied such algorithm, noise reduction won't work.

Table Of Contents

Conclusions

- We focus on low-resource scenario by MaTrEx: 4 new techniques.
- For the CHALLENGE Chinese–English translation task, our system achieved the top BLEU score among other systems.
- Word lattice
  - best single system for ZN–EN and TR–EN.
  - We show greater benefit for TR–EN (morphologically rich languages).
- Noise reduction
  - Under 3-12 percents of duplication, our noise reduction may not work (= If it's intentional, IWSLT orgnizer has more effective algorithm than ours).
- System combination techniques
  - For ZN–EN and TR–EN, the best performance is achieved.
  - Only for EN–ZH translation, slightly inferior performance.

Thank you.

Acknowledgement:

- This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation at Dublin City University.
- Irish Centre for High-End Computing.
- Kemal Oflazer for providing us the output of morphological analyser

Reference

- Okita, T., *Data Cleaning for Word Alignment*, ACL-IJCNLP Student Research Workshop, 2009.
- Okita, T., *Preprocessing Methods for Word Alignment*, CLUKI, 2009.
- H. Hassan, Y. Ma, and A. Way, *"MaTrEx: the DCU Machine Translation system for IWSLT 2007, Proceedings of the International Workshop on Spoken Language Translation*, Trento, Italy, 2007, pp. 21–28.
- Y. Ma, J. Tinsley, H. Hassan, J. Du, and A. Way, "Exploiting alignment techniques in MaTrEx: the DCU Machine Translation system for IWSLT08," in *Proceedings of International Workshop on Spoken Language Translation (IWSLT08)*, Honolulu, HI, 2008, pp. 26–33.

Reference

- Y. Ma and A. Way, "Bilingually motivated domain-adapted word segmentation for Statistical Machine Translation," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, Athens, Greece, 2009, pp. 549–557.
- ——, "Bilingually motivated word segmentation for Statistical Machine Translation," *ACM Transactions on Asian Language Information Processing, Special Issue on Machine Translation of Asian Languages*, vol. 8, no. 2, pp. 1–24, 2009.
- J. Du, Y. He, S. Penkale, and A. Way, "MaTrEx: The DCU MT system for WMT 2009," in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, 2009, pp. 95–99.

Reference

- Koehn, P., Axelrod, A., Birch, A., Callison-Burch, C., Osborne, M., Talbot, D., *Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation.*, International Workshop on Spoken Language Translation, 2005.
- J. Xu, E. Matusov, R. Zens, and H. Ney, "Integrated Chinese word segmentation in Statistical Machine Translation," in *Proceedings of the International Workshop on Spoken Language Translation*, Pittsburgh, PA, 2005, pp. 141–147.
- C. Dyer, S. Muresan, and P. Resnik, "Generalizing word lattice translation," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, OH, 2008, pp. 1012–1020.

Reference

- P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: Open source toolkit for Statistical Machine Translation,” in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, 2007, pp. 177–180.
- Z. Li, C. Callison-Burch, C. Dyer, S. Khudanpur, L. Schwartz, W. Thornton, J. Weese, and O. Zaidan, “Joshua: An open source toolkit for parsing-based machine translation,” in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, 2009, pp. 135–139.
- A. Zollmann and A. Venugopal, “Syntax augmented Machine Translation via chart parsing,” in *Proceedings of the Workshop on Statistical Machine Translation*, New York City, NY, 2006, pp. 138–141.