MaTrEx: DCU Machine Translation System for IWSLT 2007

Hany Hassan hhasan@computing.dcu.ie

Yanjun Ma yma@computing.dcu.ie

Andy Way away@computing.dcu.ie

Overview

1. DCU at IWSLT-07
2. System Description
3. Experiments & Results
4. Concluding Remarks
5. Future Work

DCU at IWSLT 2007

Second Participation following IWSLT-06:

- Classical & Challenge Tasks, in three directions:
  - Arabic $\rightarrow$ English
  - Chinese $\rightarrow$ English
  - Japanese $\rightarrow$ English
- For J2E and A2E, translated both the single-best ASR hypotheses and the correct recognition results;
- For C2E, just translated the correct recognition results.

DCU at IWSLT 2007

Three main research contributions:

- we try out our word packing technique [Ma et al., ACL-07] on different language pairs;
- we smooth t-tables with out-of-domain word translations for the A2E and C2E tasks to overcome problem of high number of OOV items;
- we deploy a translation-based model for case and punctuation restoration.

Overview

1. DCU at IWSLT-07
2. System Description
3. Experiments & Results
4. Concluding Remarks
5. Future Work

System Description

MATREX: A Hybrid EBMT/SMT System, consisting of:

- A word alignment component (GIZA++)
- A chunking component
- A chunk alignment component

System Description

- Two phrase alignment components:
  - "SMT"-style phrase aligner (standard phrase extraction from GIZA++ alignments)
  - "EBMT"-style phrase aligner ('Marker'-style phrases are extracted from (i) the chunker and (ii) the chunk aligner)
- A minimum-error rate training component (MOSES)
- A decoder (MOSES)
- A case and punctuation restoration component

Word Packing: Motivation

- Tokenization is defined in a *monolingual* context;
- Alignment is a *bilingual* task.

Monolingual segmentations do not necessarily match, especially for non-related languages like Chinese–English:

白葡萄酒: white wine	closest: 最 近
百货公司: department store	fifteen: 十 五
抱歉: excuse me	fine: 很 好
报警: call the police	flight: 次 航班
杯: cup of	get: 拿 到
必须: have to	here: 在 这里

Word Packing: Our Approach

Adapt the segmentation to the (bilingual) alignment task.

How? By 'packing' several words together when they correspond to a single word in the other language.

Packed words can be extracted from 1-to-*n* alignments (i.e. packing can rely on word alignment).

By packing words together, segmentations are made more comparable, so alignment is simplified.

Word Packing: Our Approach

1. Extract 1-to-n alignments from word alignment (candidate extraction)
2. Estimate the reliability of these candidates for packing
3. Replace the reliable candidates with a single token (word packing)
4. Re-iterate word alignment (and possibly go back to 1.)

Word Packing: Candidate Extraction

- We use the asymmetric IBM model 4 (GIZA++ implementation) to get 1-to-n alignments in both directions

- We obtain alignments $a = \langle c, E \rangle$, and $a' = \langle e, C \rangle$, where c and e are Chinese and English words, and E and C are English and Chinese consecutive sequences of words

For each alignment $a = \langle c, E \rangle$, we estimate the reliability of the candidates extracted and keep only those within some (tuned) thresholds $COOC_t$ and AC_t.

Word Packing: Iteration

- The reliable candidates are stored in 2 bilingual dictionaries (one in each direction)

- The training data are updated: for a member $a = \langle c, E \rangle$ of the dictionaries, if c and E are in a pair of aligned sentences, the sequence of words E is replaced with a single token

Starting with an initial segmentation, we:

- We can apply the method several times (parameter k);

- We stop when (i) there is no more packing to perform, or (ii) no improvement is seen on the development set.

Smoothing Translation Tables

The OOV ratio for the Arabic to English IWSLT06 test set was over 24%.

Well known that adding data from another domain degrades translation accuracy.

However, rather than combining *all* out-of-domain data with domain-specific data as is, or giving the latter a higher weight, we smooth the in-domain t-tables with lexical probabilities from out-of-domain data.

i.e. we add phrases of length one from the out-of-domain data to our in-domain phrase tables.

Smoothing Translation Tables

After performing standard bidirectional word alignment, we combine the resulting word-based t-table with the in-domain t-table to construct a larger smoothed t-table.

Note that in our experiments, using the OOV translation tables helps in translating both OOV *and* in-vocabulary items ...

For phrase translation, if we can use the in-domain phrase table we use it, otherwise we back-off to the more reliable word to word translation from the out-of-domain data.

Smoothing Translation Tables: Results

	OOV Ratio
No smoothing	24.23%
Smoothing	6.42%

Smoothing effect on OOV ratio for IWSLT-06 data

- On IWSLT-06 test set, A2E improved from 23.68 to 25.97 BLEU score, a relative improvement of 9.6%.

- On IWSLT-07 test set, C2E improved from 30.00 to 30.53 BLEU score, a relative improvement of 1.8%.

Case & Punctuation Restoration

Obviously an important task in speech translation ...

Punctuation marks: hidden events occurring between words. Find most likely hidden tag sequence (consistent with the given word sequence) via an *n*-gram LM trained on a punctuated text.

Case restoration: a disambiguation task where we choose between the (case) variants of each word of a sentence. Again, find the most likely sequence via an *n*-gram LM trained on a case-sensitive text.

Case & Punctuation Restoration

In our experiments, we consider case and punctuation restoration as a translation process:

- the case-sensitive text with punctuation can be considered as the target language;
- we remove the punctuation and case information in the target language and use them as the corresponding source language to construct a pseudo-'bilingual' corpus;
- train system to restore punctuation and/or case information.

Overview

1. DCU at IWSLT-07
2. System Description
3. Experiments & Results
4. Concluding Remarks
5. Future Work

Experiments & Results

- *Data*: used the provided BTEC datasets;
- *Training*: used default set, plus devsets 1–3 for all three language pairs;
- *Development*: used devset4;
- *Language Models*: used SRILM toolkit on target side of the bilingual training data;
- *Preprocessing*: English sentences tokenized using the MaxEnt-based tokenizer of the OpenNLP toolkit, with case information removed;
- *OOV items*: used LDC parallel news data and a large part of the UN data (2 million sentences, about 50M words);
- *Arabic data*: tokenized and segmented using the ASVM toolkit.

Experiments & Results: Arabic

Official Result:

Data Condition	BLEU
ASR output (1-best)	0.3942
Correct Transcripts	0.4709

Experiments & Results: Arabic IWSLT-06 Data

System	BLEU
Baseline	0.2253
WordPacking (WP)	0.2264

Word Packing helps a little ...

Experiments & Results: Arabic IWSLT-06 Data

System	BLEU
Baseline	0.2253
WordPacking (WP)	0.2264
WP+ Case/Punct Restoration (CP)	0.2368

Word Packing helps a little ...

Adding Case/Punctuation Restoration helps quite a bit ...

Experiments & Results: Arabic IWSLT-06 Data

System	BLEU
Baseline	0.2253
WordPacking (WP)	0.2264
WP + Case/Punct Restoration (CP)	0.2368
WP + CP + Smoothing for OOV	0.2453

Word Packing helps a little ...

Adding Case/Punctuation Restoration helps quite a bit ...

OOV Lexical Smoothing useful too ...

Experiments & Results: Arabic IWSLT-06 Data

System	BLEU
Baseline	0.2253
WordPacking (WP)	0.2264
WP + Case/Punct Restoration (CP)	0.2368
WP + CP + Smoothing for OOV	0.2453
WP + CP + Smoothing ALL	0.2597

Word Packing helps a little ...

Adding Case/Punctuation Restoration helps quite a bit ...

OOV Lexical Smoothing useful too ...

Doing Lexical Smoothing on *all* words helps even more ...

Overall, 3.4 BLEU points, or 15% relative improvement over baseline.

Experiments & Results: Japanese

Official Result:

Data Condition	BLEU
ASR output (1-best)	0.3182
Corrected Transcripts	0.3959

Experiments & Results: Japanese

Data Condition	BLEU
ASR output (1-best)	0.3182
Corrected Transcripts	0.3959
Fixing Tokenisation ASR output	0.3523

After fixing tokenisation differences between the output from MaTrEx and the reference translations provided:

- obtained a 10.7% relative improvement in BLEU score on ASR task;

Experiments & Results: Japanese

Data Condition	BLEU
ASR output (1-best)	0.3182
Corrected Transcripts	0.3959
Fixing Tokenisation ASR output	0.3523
Fixing Tokenisation Corrected Transcripts	0.4216

After fixing tokenisation differences between the output from MaTrEx and the reference translations provided:

- obtained a 10.7% relative improvement in BLEU score on ASR task;
- obtained a 6.5% relative improvement in BLEU score on Corrected Transcripts task.

Experiments & Results: Chinese

Official Result:

Data Condition	BLEU
Corrected Transcripts	0.2737

Experiments & Results: Chinese

Data Condition	BLEU
Corrected Transcripts	0.2737
Fixing Tokenisation	0.3000

Fixing tokenisation differences between MaTrEx output and references provided, 8.77% relative improvement in BLEU score.

Experiments & Results: Chinese

Data Condition	BLEU
Corrected Transcripts	0.2737
Fixing Tokenisation	0.3000
OOV Lexical Smoothing	0.3053

Fixing tokenisation differences between MaTrEx output and references provided, 8.77% relative improvement in BLEU score.

Adding OOV lexical smoothing leads to a further 1.77% increase.

Experiments & Results: Chinese

Data Condition	BLEU
Corrected Transcripts	0.2737
Fixing Tokenisation	0.3000
OOV Lexical Smoothing	0.3053
Lower Casing	0.3203

Fixing tokenisation differences between MaTrEx output and references provided, 8.77% relative improvement in BLEU score.

Adding OOV lexical smoothing leads to a further 1.77% increase.

Lower casing the output gives a 17% higher score overall than the official submitted system.

Experiments & Results: Chinese Translation Examples

ZH: 在 巴黎 出 了 交通 事故 。

baseline: in paris |0-1| out |2-3| a traffic accident |4-5| . |6-6|

WP: in paris |0-1| **there 's** |2-2| a traffic accident . |3-6|

ZH: 到 洛杉矶 需要 多 长 时间 ？

baseline: how long |3-5| do i need |2-2| to los angeles |0-1| ? |6-6|

WP: how long **will it take** |3-5| to |2-2| **get to** los angeles |0-1| ? |6-6|

Overview

1. DCU at IWSLT-07
2. System Description
3. Experiments & Results
4. Concluding Remarks
5. Future Work

Concluding Remarks

- Introduced the DCU MaTrEx Data-Driven MT system;
- Participated in the Classical & Challenge Tasks, for Chinese-, Japanese- and Arabic-to-English directions;
- Presented our novel research in:
  - Word Packing ('bilingual segmentation');
  - Smoothing t-tables with OOV lexical translations;
  - Case & Punctuation Restoration as a pseudo-MT task.

Overview

1. DCU at IWSLT-07
2. System Description
3. Experiments & Results
4. Concluding Remarks
5. Future Work

Ongoing and Future Work

- Incorporate our Supertagging target LMs [Hassan et al., ACL-07];
- Incorporate our source language context-informed features [Stroppa et al., TMI-07];
- Investigate non-contiguous word packing;
- . . .

Questions

Thank you for your attention

`http://www.nclt.dcu.ie/mt/`

This work was supported by Science Foundation Ireland (grant no. OS/IN/1732).