中科院计算所

INSTITUTE OF COMPUTING TECHNOLOGY ,CAS

An EBMT System Based on Word Alignment

HOU Hongxu, DENG Dan, ZOU Gang, YU Hongkui, LIU Yang, XIONG Deyi and LIU Qun

hxhou@ict.ac.cn

Introduction

- EBMT
- Based on word alignment
- A 220K sentence pairs' corpus
- 5 steps of translation

System Architecture

- Corpus
  - 220k sentence pairs
  - 460k words and phrases
- Program
  - Matching and searching
  - Fragment matching
  - Fragment assembling
  - Evaluation
  - Post processing

Corpus

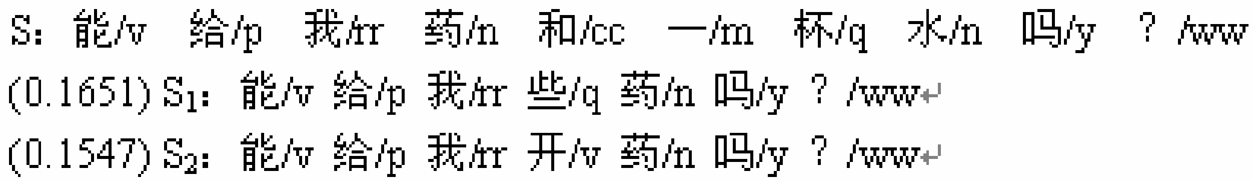
- Scale
  - 220K sentence pairs, includes news, literal, dictionaries and dialogues
  - 460K words and phrases dictionary
- POS tagged/tokenized
  - ICTCLAS 2.0
  - Accuracy: 97.8%
- Word aligned
  - Based on large dictionary
  - Precision: 84.0%
  - Recall: 62.9%

Matching and Searching

- Searching the most similar sentences
- Measure of similarity
  - Weight of POS
  - Run length

Matching and Searching (cont.)

- *An example*



S：能/v 给/p 我/rr 药/n 和/cc 一/m 杯/q 水/n 吗/y ？/ww

(0.1651) S_1：能/v 给/p 我/rr 些/q 药/n 吗/y ？/ww↵

(0.1547) S_2：能/v 给/p 我/rr 开/v 药/n 吗/y ？/ww↵

Matching and Searching (cont.)

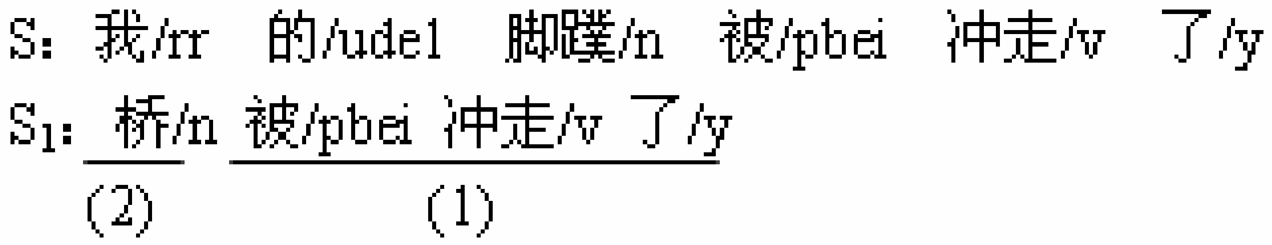
- Searching
  - A index of words/POS
  - Remove high frequency words.
  - Remove named entities
  - Short sentences

Fragment matching

- Fragment matching
  - Matching and mismatching fragment
  - Word matched and POS matched

Fragment matching (cont.)

- *An example*



S： 我/rr 的/ude1 脚蹼/n 被/pbei 冲走/v 了/y

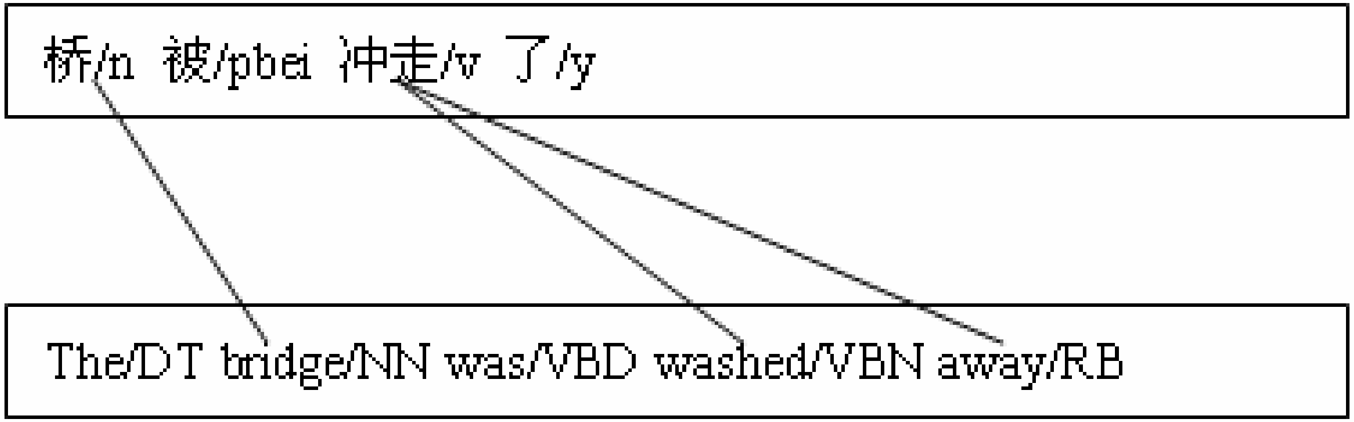
S_1: $\underbrace{\text{桥/n}}_{(2)}$ $\underbrace{\text{被/pbei 冲走/v 了/y}}_{(1)}$

Target fragment finding

- Based on word alignment
  - Target contents
    - Which words should be placed
  - Target positions
    - Which position should be placed

Target fragment finding

- *An example*

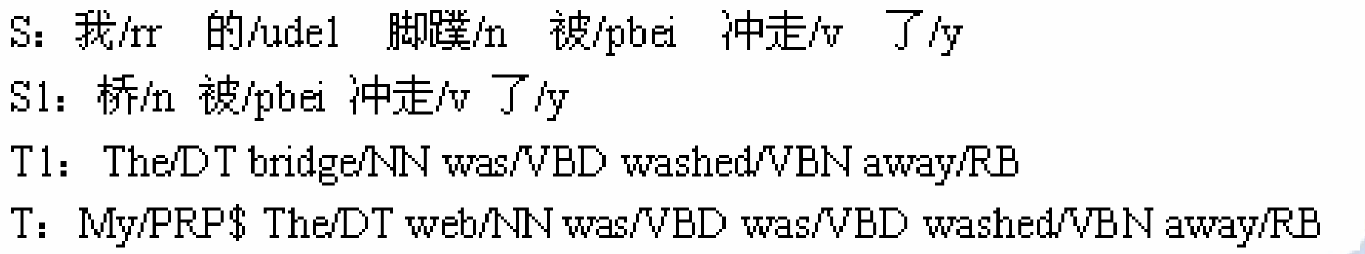


桥/n 被/pbei 冲走/v 了/y

The/DT bridge/NN was/VBD washed/VBN away/RB

Fragment assembling

- *An example*



S：我/rr　的/ude1　脚蹼/n　被/pbei　冲走/v　了/y

S1：桥/n 被/pbei 冲走/v 了/y

T1：The/DT bridge/NN was/VBD washed/VBN away/RB

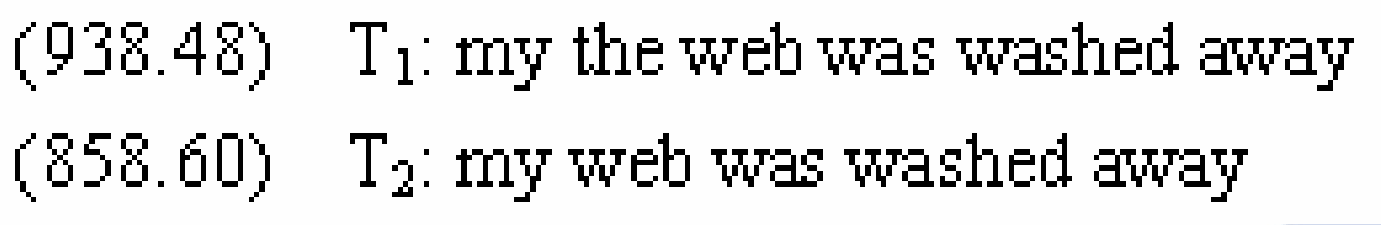
T：My/PRP$ The/DT web/NN was/VBD was/VBD washed/VBN away/RB

Make choices

- Whether the non-aligned words should be appeared in the target sentence
- N-gram

Make choices (cont.)

- *An example*



(938.48) T_1: my the web was washed away

(858.60) T_2: my web was washed away

Evaluation results

	Official results	Correct results
BLEU	0.0798 (9)	0.2013 (8)
GTM	0.3862 (9)	0.6380 (5)
NIST	3.6443 (9)	6.4716 (5)
WER	0.8466 (9)	0.6275 (8)
PER	0.7650 (9)	0.5187 (7)
Fluency	2.7180 (7)	
Adequacy	3.0820 (5)	

* We have made a mistake in submitting result

Some cases

- Existed in corpus
- Good fragments of an example
- Replace some words
- Part of example
- Combine of phrases
- Small fragments

Further improvement

- Phrase recognition
- Word alignment
- Word cluster
- Sentence classify
- Corpus

Thank you