The *ICT 's Patent MT System Description for NTCIR-9*

Hao Xiong, Linfeng Song, Fandong Meng, Yajuan Lü, Qun Liu
Key Lab. of Intelligent Information Processing
{xionghao, songlinfeng, mengfandong, lvyajuan, liuqun}ict.ac.cn

Overview

We participate three subtasks, and submit 6 translation results for each subtask

- **Chinese-English: 6 system submissions**
- **English-Japanese: 6 system submission**
- **Japanese-English: 6 system submissions**

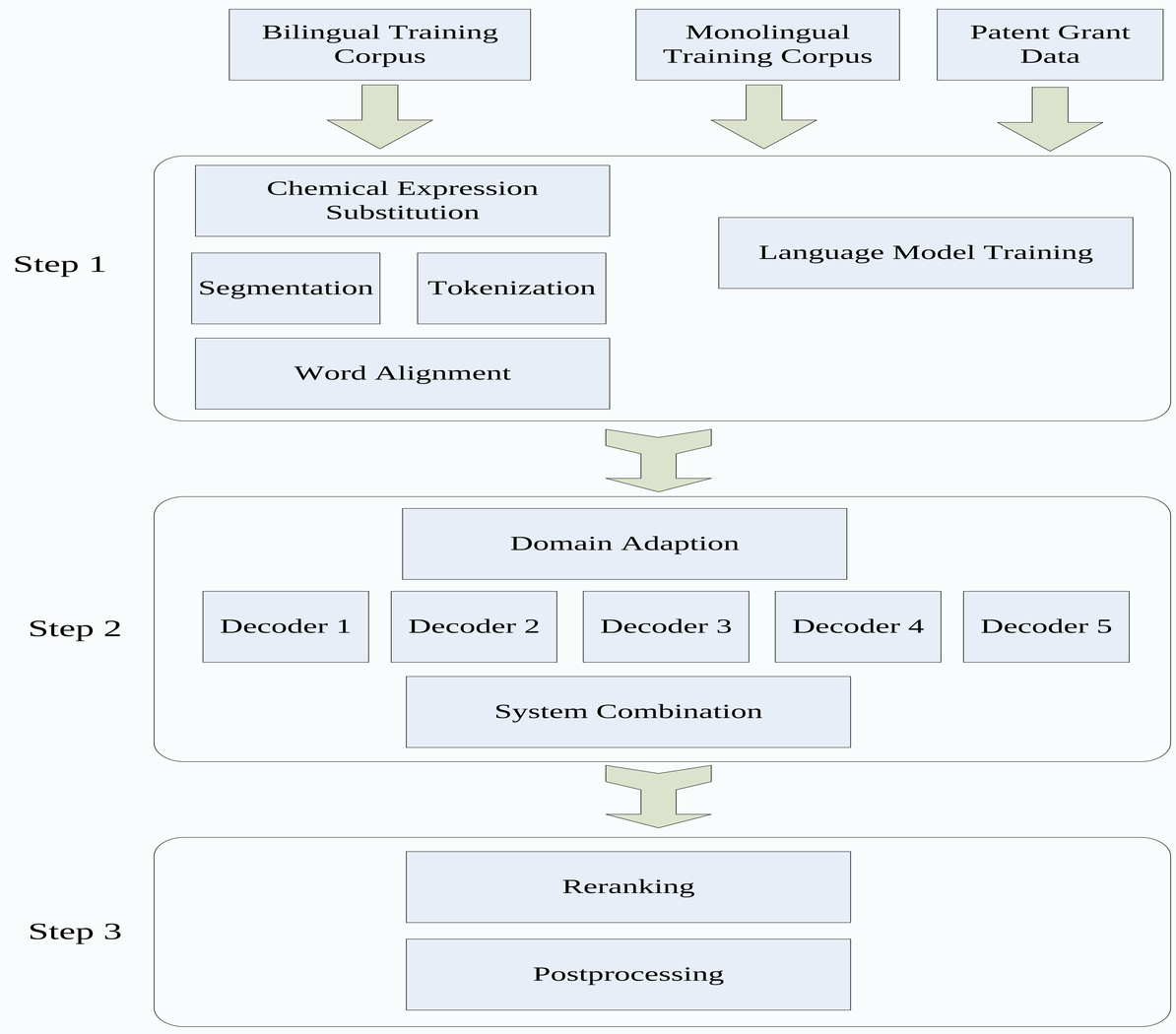
We employ three types of decoders

- **In-house implemented Hierarchical phrase-based model (HPB)**
- **Phrase-based model (Moses)**
- **Combinational system (SCM)**

Main techniques

- **Chemical Expression Substitution**
- **Refined Segmentation**
- **Domain Adaption**
- **Multi-system Reranking**

Overall Architecture



Chemical Expression Substitution

Sample: "向多颈烧瓶中投入1,6-二氢-6-氧代-4-嘧啶羧酸" => "向多颈烧瓶中投入 CHEM1"

1. **First identify three location words "二氢", "氧代" and "嘧啶羧酸" using location dictionary**
2. **Since the character after the first "-" is "二", two Arabic numbers before it are recognized as parts of this chemical expression**
3. **With the same procedure, we will recognize the whole chemical expression "1,6-二氢-6-氧代-4-嘧啶羧酸"**

Refined Segmentation

1. **Collecting a terminological dictionary**
2. **For a given sentence, we find out all the words which are also in the dictionary, and store them in a queue in order.**
3. **When a word is in the dictionary queue, it will be re-weighted during decoding procedure.**

Domain Adaption

1. **Classify training corpus into identical domains.**
2. **Classify developing corpus into identical domains.**
3. **Tune weights independently for each domain.**
4. **Classify testing corpus into identical domains.**
5. **Translate testing sentence using corpus from its corresponding domain.**

Multi-system Reranking

1. **Bootstrap N new development sets from the original set**
2. **Tune a subsystem using each newly generated set**
3. **For each sentence, generate a k-best candidate list from each subsystem**
4. **Rerank the integrated k-best list by the sum of voting score from each subsystem**

Data Usage

System	Bilingual	Monolingual
C-E	1 Million	40 Million
J-E	3 Million	40 Million
E-J	3 Million	73 Million

Baseline Systems

System	C-E	J-E	E-J
HPB	30.08	23.55	32.85
Moses	29.56	23.31	32.27
SCM	**30.73**	**25.29**	**33.50**

Language Model Experiments

System	C-E	J-E	E-J
HPB	31.28	25.27	34.04
Moses	30.78	25.28	33.72
SCM	**32.67**	**25.63**	**34.55**

Rule Filter Experiments

System	C-E	J-E	E-J
P=0.0	31.28	25.27	**34.04**
P=0.9	**31.72**	25.66	33.45
P=1.0	31.67	25.46	33.42
P=1.1	31.51	**25.86**	33.51

Incorporating templates(not submit)

System	C-E
Baseline	**30.08**
Template-based	29.39

Chemical Expression Substitution(not submit)

System	C-E
Baseline	30.08
CES	**31.19**

Final Results

- ***sys1*: combinational system using following systems.**
- ***sys2*: HPB with rule filter threshold p = 1.1.**
- ***sys3*: HPB with rule filter threshold p = 1.0.**
- ***sys4*: HPB with rule filter threshold p = 0.9.**
- ***sys5*: HPB with rule filter threshold p = 0.0.**
- ***sys6*: Moses**

System	C-E	J-E	E-J
sys1	**31.97**	**27.28**	**32.91**
sys2	31.52	26.90	32.10
sys3	31.57	26.55	31.72
sys4	30.78	26.71	32.06
sys5	30.76	26.06	32.17
sys6	30.64	26.84	30.17

Refined Segmentation(not submit)

System	C-E
Baseline	**30.08**
Refined Segmentation	29.39

Domain Adaption(not submit)

System	C-E
Baseline	30.08
Domain Adaption	**31.19**

Multi-document Reranking(not submit)

System	C-E
Baseline	31.08
reranking	**31.90**