Hierarchical system combination for machine translation

Author：Fei Huang, Kishore Papineni

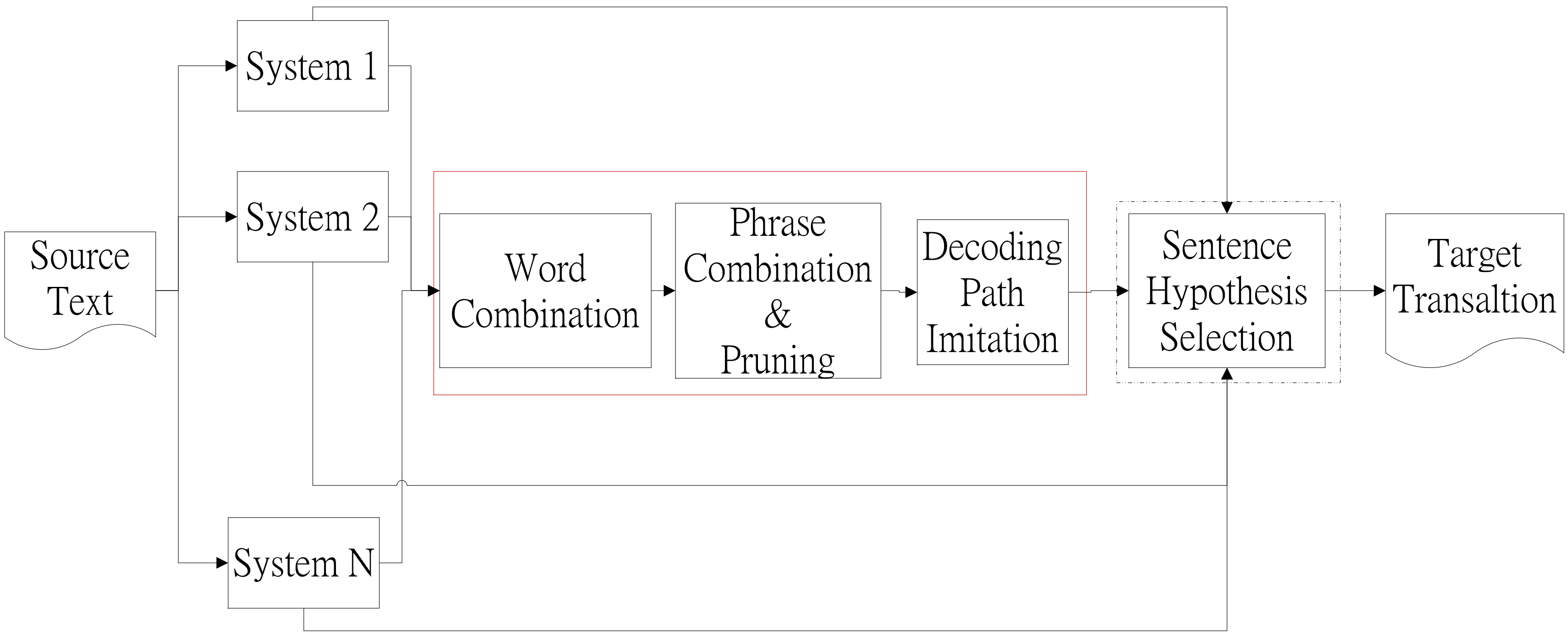
Professor：陳嘉平

Reporter：陳逸昌

Introduction

- System combination has been conducted in two ways
  - Glass-box
  - Black-box
- In this paper, they introduce the hierarchical system combination strategy
  - This approach allows combination on word, phrase and sentence levels

Hierarchical system combination framework



Baseline MT system

- Select an MT system to retranslate the test sentences with the refined models
- The decoder tries to find the translation hypothesis with the minimum translation cost
- The overall cost is the log-linear combination of different feature function

Hierarchical system combination framework

- Each system produces both top-one translation hypothesis through a XML file format
  - Source sentence is segmented into a sequence of phrases
  - The order and translation of each source phrases
  - A vector of feature scores for the whole test sentence

XML format


```
<tr engine="XXX">
<s id="0"> <w> اردوغان </w><w> يؤكد </w><w> بان </w><w> تركيا </w><w> سترفض
</w><w> اي </w><w> ضغوطات </w><w> لحثها </w><w> علي </w><w> الاعتراف </w><w>
بقبرص </w></s>
<hyp r="0" c="2.15357">
 <t>
<p al="0-0" cost="0.0603734"> erdogan </p>
<p al="1-1" cost="0.367276"> emphasized </p>
<p al="2-2" cost="0.128066"> that </p>
<p al="3-3" cost="0.0179338"> turkey </p>
<p al="4-5" cost="0.379862"> would reject any </p>
<p al="6-6" cost="0.221536"> pressure </p>
<p al="7-7" cost="0.228264"> to urge them </p>
<p al="8-8" cost="0.132242"> to</p>
<p al="9-9" cost="0.113983"> recognize </p>
<p al="10-10" cost="0.133359"> Cyprus </p>
</t>
<sco>
19.6796 8.40107 0.333514 0.00568583 0.223554 0 0.352681 0.01 -0.616 0.009 0.182052
</sco>
</hyp>
</tr>
```

Word combination

- The goal is to construct a testset-specific word translation model

- Word alignments are identified within a phrase pair based on IBM model 1

- Collect word alignment counts from the whole test set translation and estimate both direction translation probability

Word combination

$$t''\left(e\middle|f\right)=\gamma t'\left(e\middle|f\right)+(1-\gamma)t\left(e\middle|f\right)$$

- $t'\left(e\middle|f\right)$ is the testset-specific source-to-target word translation probability
- $t\left(e\middle|f\right)$ is the probability from general model
- γ is the linear combination weight in this paper = 0.8

Phrase Translation Combination and Pruning

- Collect and merge phrase translation tables from each system

$$P'(e|f) = \frac{C_b(f,e) + \sum \alpha_m C_m(f,e)}{C_b(f) + \sum \alpha_m C_m(f)}$$

- C_b is the phrase pair count from the baseline decoder
- C_m is the count from other systems
- α_m is a system-specific linear combination weight

Phrase Translation Combination and Pruning

- The corresponding phrase translation cost is updated as $S'(e,f) = -\log P'(e,f)$

- Another phrase combination strategy works on the sentence level
- It collects phrase translation pairs used by different MT systems to translate the same sentence

$$S''(e|f) = \frac{\beta}{|C(f,e)|} \times S'(e,f)$$

Phrase Translation Combination and Pruning

- The combine phrase table contains multiple translation for each source phrase
- Many of them are unlikely translation given the context
- Only keep phrase pairs whose target phrase is covered by existing system translation

Decoding Path Imitation

- A reordering cost function that encourages search along decoding paths adopted by other decoders
- Specifically, give a partially expanded path $P = \{s_1 < s_2 < \cdots < s_m\}$, word pair $(s_i < s_j)$ is covered by a full decoding path Q (from other system outputs), we denote the relationship as $(s_i < s_j) \in Q$

Decoding Path Imitation

- For any ordered word pair $\left(s_i < s_j\right) \in P$, denote its matching ratio as the percentage of full decoding paths that cover it

$$R\left(si < sj\right) = \frac{|Q|}{N}, \left\{Q \middle| \left(s_i < s_j\right) \in Q\right\}$$

- Path matching cost function

$$L(P) = -\log \frac{\sum_{\forall\left(s_i < s_j\right) \in P} R\left(s_i < s_j\right)}{\sum_{\forall\left(s_i < s_j\right) \in P} 1}$$

- Path P : 1 < 2 < 3 < 4
- System 1 : 1 < 2 < 4 < 3
- System 2 : 1 < 3 < 2 < 4

$$R(1<2)=1 \quad R(1<3)=1 \quad R(1<4)=1$$

$$R(2<3)=\frac{1}{2} \quad R(2<4)=1$$

$$R(3<4)=\frac{1}{2}$$

$$L(P)=-\log\left(\frac{6}{7}\right)$$

Sentence Hypothesis Selection

- The sentence hypothesis selection module only take the final translation outputs from individual systems, including the output from the combination
- Typical 5-gram word language model

$$\begin{aligned} E' &= \arg\min\left(-\log P_{5glm}\left(E\right)\right) \\ &= \arg\min\sum_i -\log p\left(e_i \middle| e_{i-4}^{i-1}\right) \end{aligned}$$

Sentence Hypothesis Selection

- Another feature function is based on the 5-gram LM score calculated on the mixed stream of word and POS tags of the translation output

- Keep the word identities of top N frequent words (N=1000 in the paper)

- Remaining words are replaced with their POS tags

Original Sentence:

in *short* , making a good plan at the *beginning* of the construction is the *crucial measure* for *reducing haphazard* economic development .

Word-POS mixed stream:

in JJ , making a good plan at the NN of the construction is the JJ NN for VBG JJ economic development .

Sentence Hypothesis Selection

$$E^* = \arg\min_E - \log P_{wplm}(E)$$

$$= \arg\min_E \sum_i - \log p\left(T(e_i) \middle| T(e)_{i-4}^{i-1}\right)$$

$$T(e) = e \text{ when } e \leq N$$

$$T(e) = POS(e) \text{ when } e > N$$

Experiments

- NIST 2003 Arabic-English
- Include 260K sentence pairs, 10.8M Arabic words and 13.5M English words
- Report results using BLEU and TER

Phrase translation combination

	BLEUr4n4c	TER
sys1	0.5323	43.11
sys4	0.4742	46.35
Tstcom	0.5429	42.64
Tstcom+Sentcom	0.5466	42.32
Tstcom+Sentcom+Prune	0.5505	42.21

$$P'(e|f) = \frac{C(f,e) + \sum \alpha_m C_m(f,e)}{C(f) + \sum \alpha_m C_m(f)}$$

- Tstcom: $S'(e, f) = -\log P'(e, f)$
- Setcom: $S''(e|f) = \frac{\beta}{|C(f,e)|} \times S'(e, f)$

Word translation combination

	BLEUr4n4c	TER
sys1	0.5323	43.11
sys2	0.5320	43.06
SentSel-word:	0.5354	42.56
SentSel-wpmix:	0.5380	43.06

- Word: typical 5-gram language model(2.9G words)
- Wpmix: word-POS mixed language model(136M words)

	BLEUr4n4c	TER
sys1	0.5323	43.11
sys2	0.5320	43.06
sys3	0.4922	46.03
sys4	0.4742	46.35
WdCom	0.5339	42.60
WdCom+PhrCom	0.5528	41.98
WdCom+PhrCom+Path	0.5543	41.75
WdCom+PhrCom+Path+SenSel	0.5565	41.59