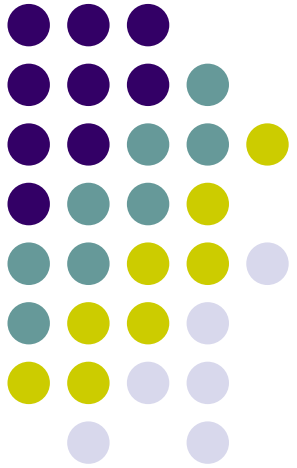
Improved Monolingual Hypothesis Alignment for Machine Translation System Combination

Author：Xiaodong He et.al

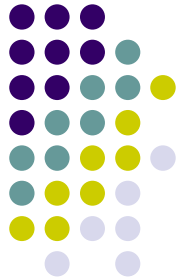
Source: ACM Transactions on Asian Language Information Processing (TALIP) Volume 8， Issue 2 , 2009

Professor：陳嘉平

Reporter：陳逸昌

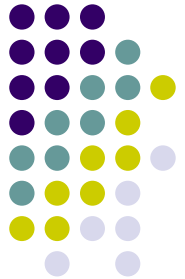


Outline



- Introduction
- Confusion-network-based MT system combination
- Indirect-HMM-based hypothesis alignment
- Experiments

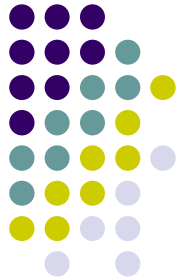
Introduction(1/2)



- High-quality hypothesis alignment is crucial to the performance of the resulting system combination

- Two challenging issues that make MT hypothesis alignment difficult
  - Synonym
  - Word order

Introduction(2/2)



- Unlike traditional HMMs whose parameters are trained via maximum likelihood estimation (MLE), the parameter of the IHMM are estimated indirectly from a variety of source including
  - Word semantic similarity
  - Word surface similarity
  - Distance-based distortion

Confusion-network-based MT system combination

E_1	he have good car
E_2	he has nice sedan
E_3	it a nice car
E_4	a sedan he has

(a) hypothesis set

$$E_B = \arg\min_{E' \in \mathbf{E}} \sum_{E \in \mathbf{E}} TER(E', E)$$

e.g., $E_B = E_1$

(b) backbone selection

E_B he have ε good car

E_4 a ε sedan he has

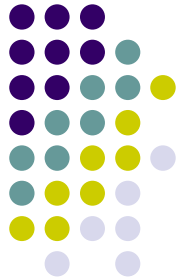
(c) hypothesis alignment

he	have	ε	good	car
he	has	ε	nice	sedan
it	ε	a	nice	car
he	has	a	ε	sedan

(d) confusion network

Figure 1: Confusion-network-based MT system combination.

Indirect-HMM-based hypothesis alignment

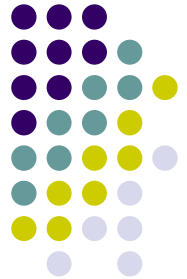


- IHMM for hypothesis
- Estimation of the similarity model
- Estimation of the distortion model
- Alignment normalization

IHMM for hypothesis(1/3)

- $e_1^I = (e_1, \ldots, e_I)$ denote the backbone
- $e_1'^J = (e_1', \ldots, e_J')$ is a hypothesis to be aligned to e_1^I
- Treat each word in the backbone as an HMM state
- Treat the words in the hypothesis as the observation sequence

IHMM for hypothesis(2/3)



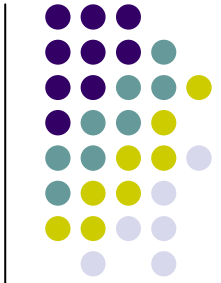
- Use a first-order HMM, assuming that
  - Emission probability $p\left(e'_j \middle| e_{a_j}\right)$ depends only on the backbone word
  - Transition probability $p\left(a_j \middle| a_{j-1}, I\right)$ depends only on the position of the last state and the length of the backbone
- Treating the alignment as hidden variable, the conditional probability that the hypothesis is generated by the backbone is given by

$$p\left(e_1'^J \middle| e_1^I\right) = \sum_{a_1^J} \prod_{j=1}^{J} \left[p\left(a_j \middle| a_{j-1}, I\right) p\left(e'_j \middle| e_{a_j}\right) \right]$$

IHMM for hypothesis(3/3)

- In their method
  - Associate a null with each backbone word to allow generating hypothesis words that do not align to any backbone word
  - Emission probabilities model the similarity between a backbone word and a hypothesis word and will be referred to as the similarity model
  - Transition probabilities model word reordering and will be called the distortion model

Estimation of the similarity model



- The similarity model
  - Models the similarity between a backbone word and a hypothesis word
  - Derived based on both semantic similarity and surface similarity

$$p\left(e'_j \middle| e_i\right) = \alpha \cdot p_{sem}\left(e'_j \middle| e_i\right) + (1-\alpha) \cdot p_{sur}\left(e'_j \middle| e_i\right)$$

- $p_{sem}\left(e'_j \middle| e_i\right)$ and $p_{sur}\left(e'_j \middle| e_i\right)$ reflect the semantic and surface similarity between e'_j and e_i
- α is the interpolation factor (in experiment α = 0.3)

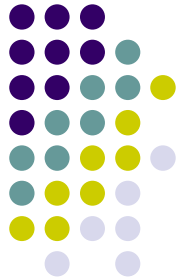
Semantic similarity model (1/3)

$$P_{sem}\left(e'_j \middle| e_i\right) = \sum_{k=0}^{K} p\left(f_k \middle| e_i\right) p\left(e'_j \middle| f_k, e_i\right)$$

$$\approx \sum_{k=0}^{K} p\left(f_k \middle| e_i\right) p\left(e'_j \middle| f_k\right)$$

- $f_1^K = \left(f_1, \ldots, f_k\right)$ is the source sentence

Semantic similarity model (2/3)



$$p(e'_j \mid f_k) = p_{s2t}\left(e'_j \mid f_k\right)$$

- $p_{s2t}\left(e'_j \mid f_k\right)$ is the translation model from the source-to-target word alignment model

Semantic similarity model (3/3)

$$p(f_k|e_i) = \frac{p_{t2s}(f_k|e_i)}{\sum_{k=0}^{K} p_{t2s}(f_k|e_i)}$$

- $p_{t2s}(f_k|e_i)$ is the translation model from the target-to-source

Surface similarity model

$$p_{sur}\left(e'_j \middle| e_i\right) = \exp\left\{\rho \cdot \left[s\left(e'_j, e_i\right) - 1\right]\right\}$$

- ρ is a smoothing factor (in experiment ρ = 3)
- $s\left(e'_j, e_i\right)$ is computed as $s\left(e'_j, e_i\right) = \dfrac{M\left(e'_j, e_i\right)}{\max\left(\left|e'_j\right|, \left|e_i\right|\right)}$
- $M\left(e'_j, e_i\right)$ is the length of the LMP (longest matched prefix) of e'_j and e_i

Estimation of the distortion model(1/4)

- Assume the transition probability $p\left(a_j = i \middle| a_{j-1} = i', I\right)$ only depend on the jump distance $\left(i - i'\right)$

$$p\left(i \middle| i', I\right) = \frac{c\left(i - i'\right)}{\sum_{l=1}^{I} c\left(l - i'\right)}$$

- They group the distortion parameters $\left\{c\left(d\right)\right\}, d = i - i$ into a few buckets

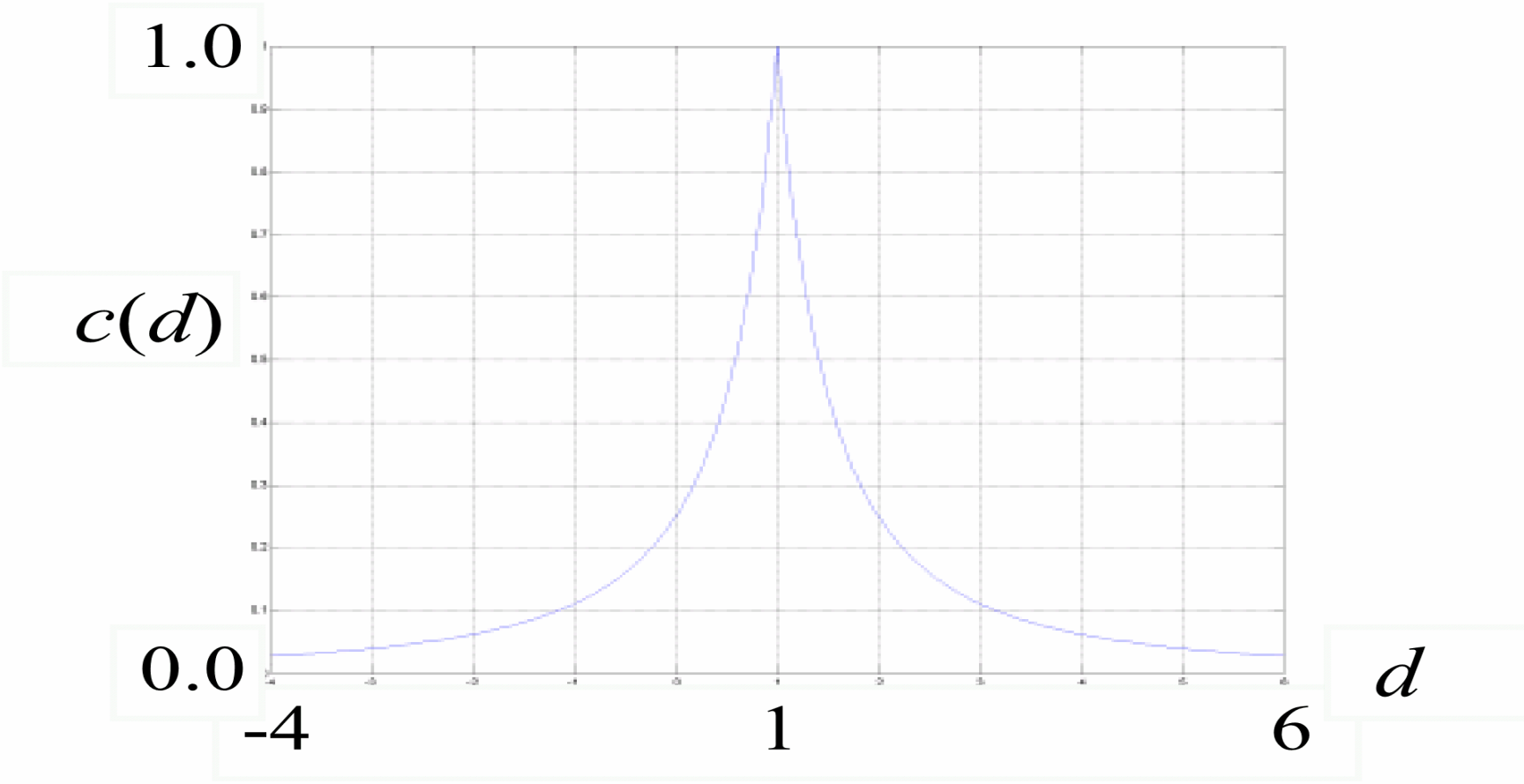
- In their implementation, 11 buckets are used for $c\left(\leq -4\right), c\left(-3\right), \ldots c\left(0\right), \ldots, c\left(5\right), c\left(\geq 6\right)$

Estimation of the distortion model(2/4)

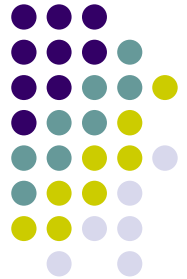
$$c(d) = \left(1 + \left|d - 1\right|\right)^{-k}, d = -4, \ldots, 6$$

- k is a tuning factor optimized on held-out data (in experiment k = 2)

Estimation of the distortion model(3/4)



Estimation of the distortion model(4/4)



- They use a fixed value p_0 for the probability of jumping to a null state

- The overall distortion model becomes

$$\tilde{p}\left(i\middle|i',I\right)=\begin{cases}p_0 & \text{if } i=null \text{ state}\\ \left(1-p_0\right)\cdot p\left(i\middle|i',I\right) & \text{otherwise}\end{cases}$$

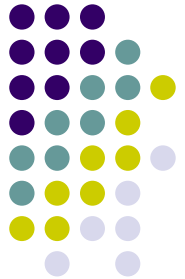
Alignment normalization(1/4)

$$\hat{a}_1^J = \arg\max_{a_1^J} \prod_{j=1}^{J} \left[p\left(a_j \middle| a_{j-1}, I\right) p\left(e'_j \middle| e_{a_j}\right) \right]$$

- The alignment produced by the algorithm can't be used directly to build confusion network

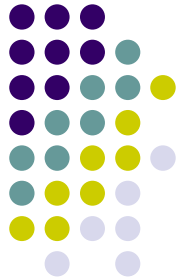
- Two reason
  - The alignment produced may contain 1-N mappings between the backbone and the hypothesis
  - If hypothesis words are aligned to a null in the backbone, they need insert actual nulls into the right places

Alignment normalization(2/4)



- Whenever more than one hypothesis words are aligned to one backbone word (1-N)
  - Keep the link which the highest probability
  - Other hypothesis words aligned to the backbone word will be aligned to the null associated with that backbone word

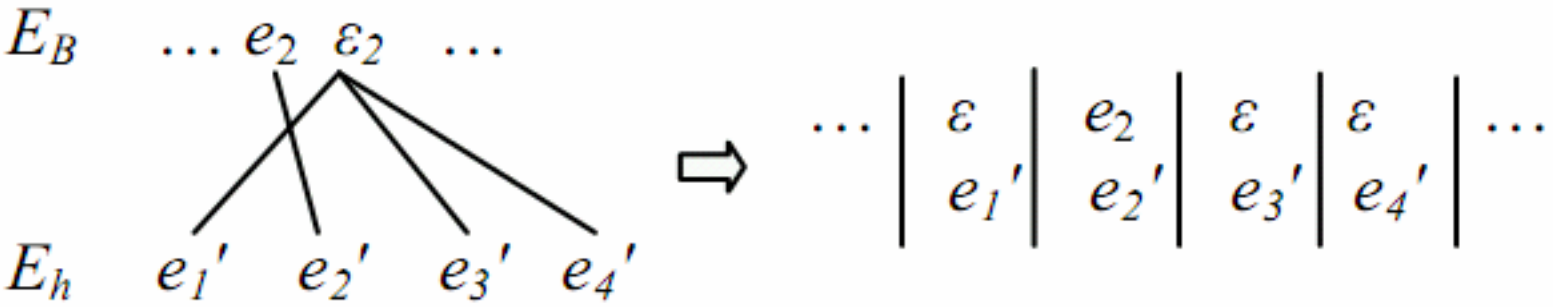
Alignment normalization(3/4)



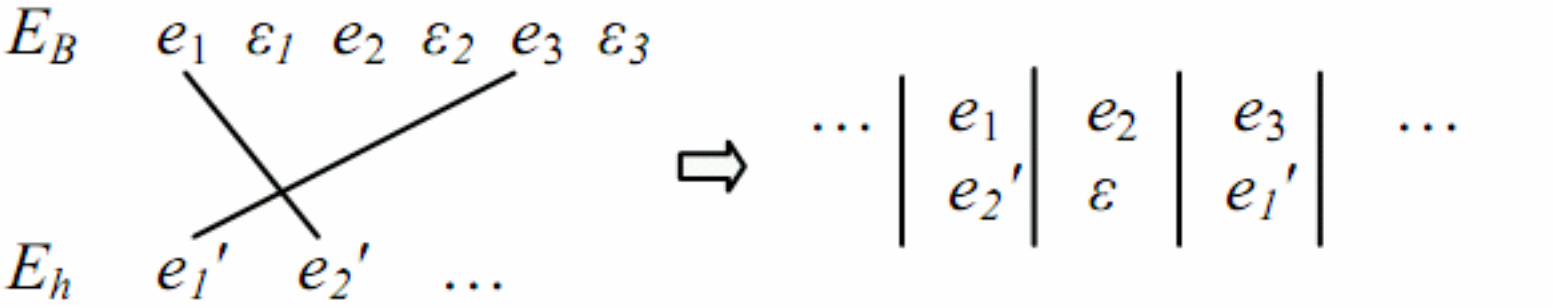
- Hypothesis words are aligned to the backbone null
  - A set of nulls are inserted around that backbone word associated with the null such that no links cross each other (case a)

- A backbone word is aligned to no hypothesis word
  - Inserted right after the hypothesis word which is aligned to the immediately preceding backbone word (case b)

Alignment normalization(1/4)

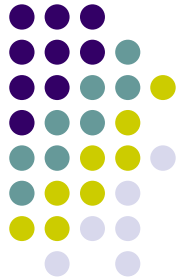


(a) hypothesis words are aligned to the backbone *null*



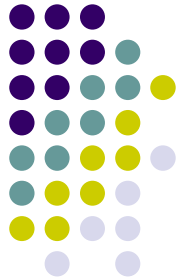
(b) a backbone word is aligned to no hypothesis word

Experiment



- Implementation details
- Development and test data
- Experimental result

Implementation details



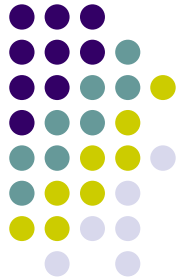
- The backbone is selected with MBR
- Each word in he confusion network is associated with a word posterior probability
- Two language model are used
  - A trigram model estimated from the English side of the parallel training data
  - A 5-gram model trained on the English GigaWord corpus from LDC

Implementation details

- The bilingual translation models trained from on two million parallel sentence-pairs selected from the training corpus of MT08

- In order to reduce the fluctuation of BLEU scores caused by the inconsistent translation output length, they compute an expected length ratio between the MT output and the source sentences on the development set

Development and test data



- The development set used for system combination parameter training contains 1002 Chinese to English sentences
  - 35% from MT04
  - 55% from MT05
  - 10% from MT06
- Test set is the MT08 Chinese to English test set, which include 1357 sentences from both newswire and Web-data

Experimental result

System	Dev ciBLEU%	MT08 BLEU%
Sys- 1	34.08	21.75
Sys-2	33.78	20.42
Sys- 3	34.75	21.69
Sys-4	37.85	25.52
Sys- 5	37.80	24.57
Sys- 6	37.28	24.40
Sys- 7	32.37	25.51
Sys- 8	34.98	26.24
TER	42.11	29.89
IHMM	43.62	30.89