Combination of Machine Translation Systems via Hypothesis Selection from Combined N-Best Lists

Author：Almut Silja Hildebrand, Stephan Vogel

Professor：陳嘉平

Reporter：陳逸昌

Outline

- Introduction
- Features
- Experiments
- Conclusions

Introduction

- Select the best hypothesis according to several feature scores
- Independent of internal translation system scores
- Possible to use non-statistical translation systems in the combination

Features

- Recalculate all feature scores in a consistent manner
  - Language models
  - Statistical word lexica
  - Position dependent n-best list word agreement
  - Position independent n-best list n-gram agreement
  - N-best list n-gram probability

Language models

- Language models with n-gram lengths of four and five
- Summing the log-probability for each word, given its history
- Normalize the sentence log-probability with the target sentence length

Statistical word lexica

$$P_{lex}(e|f_1^J) = \frac{1}{J+1}\sum_{j=0}^{J} p(e|f_j)$$

- f_1^J is the source sentence
- J is the source sentence length
- $p(e|f_j)$ is the lexicon probability of the target word e, given one source word f_j

Statistical word lexica

- The above equation is the sum of all translation probabilities of e for each word f_j from the source sentence f_1^J

- Calculate an average word translation probability as the sentence score, they sum over all word e_i in the target sentence and normalize with the target sentence length I

Maximum lexicon model

- The sum in above equation is dominated by the maximum lexicon probability
- Use the maximum lexicon probability as an additional feature

$$P_{lex-\max}(e|f_1^J) = \max_j p(e|f_j)$$

Deletion model

- If a word lexicon probability falls under a threshold, this word is considered a deletion

- The feature simply calculate the percentage of deletion words

Position dependent n-best list word agreement

$$h_k(e_i) = \frac{1}{N_k} \sum_{n=1}^{N_k} \delta(e_{n,i}, e)$$

- N_k is the number of entries in the n-best list for the corresponding source sentence k
- $e_{n,i}$ is the word at position i in the n_{th} hypothesis
- Note only the same word choice in the translation, but also the same word order

Position dependent n-best list word agreement

$$h_k(e_i) = \frac{1}{N_k} \sum_{n=1}^{N_k} \delta(e_{n,i-t} \cdots e_{n,i+t}, e)$$

- Based on a window of size $i \pm t$ around position i
- Use window sizes for t = 0 to t = 2
- The score for the target sentence is sum over all word agreement scores normalized by the sentence length

Position independent n-best list n-gram agreement

$$h_k(e_{i-(n-1)}^{i}) = \frac{1}{N_k} \sum_{j=1}^{N_k} \delta(e_{i-(n-1)}^{i}, e_{1,j}^{I})$$

- The relative frequency of target sentences in the n-best list for one source sentence, that contain that n-gram
- Use n-gram lengths = 1…6
- The score for the target sentence is sum over all word agreement scores normalized by the sentence length

N-best list n-gram probability

$$p(e_i \mid e_{i-(n-1)}^{i-1}) = \frac{C(e_{i-(n-1)}^{i})}{C\left(e_{i-(n-1)}^{i-1}\right)}$$

- The n-best list n-gram probability is a traditional n-gram language model probability
- The counts for the n-gram are collected on the n-best list entries from one source sentence only

Experiments

- Evaluation
- Models
- Systems
- Feature impact
- N-best lists size
- Combination of all systems

Evaluation

- Chinese to English
  - Development set:MT03 test set contains 919 sentences with 4 reference
  - Test sets:MT06 test set contains 1099 sentence with 4 reference
- Report results using BLEU and TER

Models

- 23 features in total
- Language model were trained from the English Gigaword Corpus V3 and the English side of the NIST training data
- Statistical word lexica were trained on the Chinese-English bilingual corpora relevant to GALE available through the LDC

Systems

- Output from six different Chinese-English translation systems.
- Trained on data for the GALE and NIST
- They are based on phrase based, hierarchical and example based translation principles, built by three translation research groups

Systems

system	MT03	MT06 BLEU	MT06 TER
A	34.68	31.45	59.43
B	35.16	31.28	57.92
C	34.98	31.25	57.55
D	34.70	31.04	57.20
E	33.50	30.36	59.32
F	28.95	26.00	62.43

Feature impact

- Ran our setup with the two language models only as a simple baseline.

- Added word lexicon features to the language model probabilities as a second baseline.

Feature impact

features	MT03	MT06 BLEU / TER
LM only	35.13	31.17 / 59.34
LM+Lex	36.96	30.97 / 59.41
no LM	39.10	32.83 / 56.23
no Lex	39.62	33.61 / 56.88
no WordAgr	39.59	33.67 / 57.25
no NgrAgr	39.45	33.47 / 56.58
no NgrProb	39.69	33.65 / 57.40
LM+NgrAgr	39.45	33.58 / 57.15
all	39.76	33.72 / 56.79

Feature impact

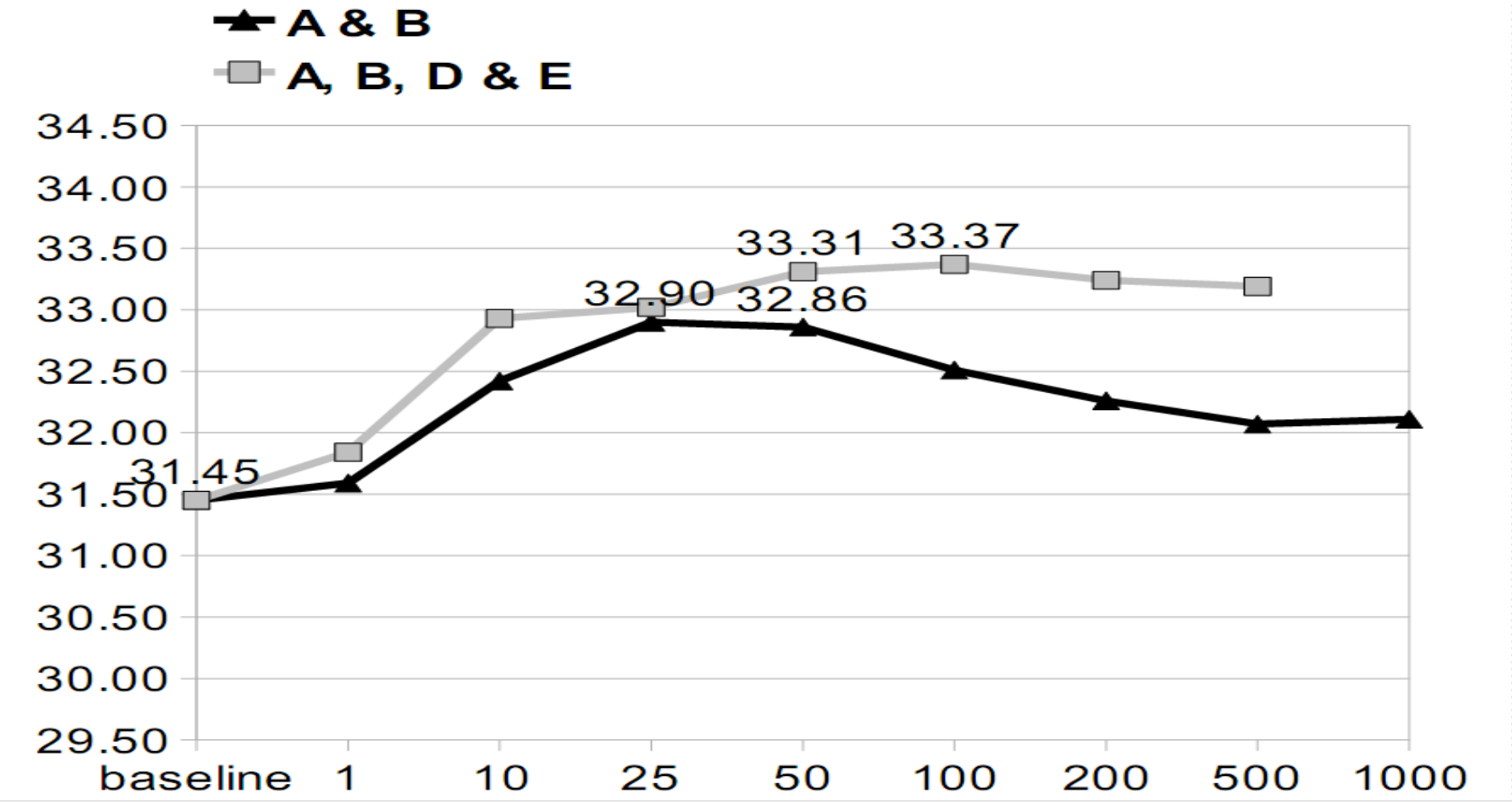
- The biggest drops are caused by removing the language model(-0.98) and the n-gram agreement(-0.25)
- Using only LM and n-gram agreement result up to 33.58 BLEU
- BUT using all features still remains the best choice

N-best list size

- Hasan et al. (2007) found that using more than 10,000 hypotheses does not help to improve the translation quality

- The difference between using 1000 and 10,000 hypotheses was very small

N-best lists size



Combination of all systems

system	baseline	combined
A	31.45	31.76 / 58.95
+ B	31.28	32.86 / 57.90
+ C	31.25	33.32 / 56.87
+ D	31.04	33.51 / 56.77
+ E	30.36	**33.72** / 56.79
+ F	26.00	33.63 / **56.45**

Conclusions

- Can be extended to use more feature scores.
- Ex:
  - Sentences length penalty
  - Source-target punctuation match
  - Phrase tables