Using Hypothesis Selection Based Features for Confusion Network MT System Combination

Sahar Ghannay and Loïc Barrault

LIUM, University of Le Mans
France

EACL 2014, Third Workshop on Hybrid Approaches to Translation

Plan

MT system combination

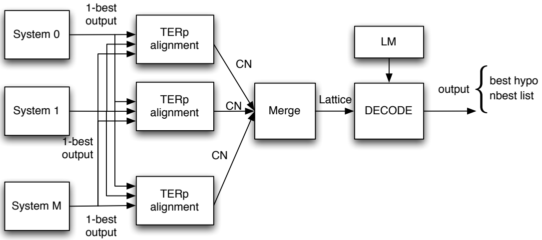
- Taken a great importance these past few years
- Resulting outputs with a better performance than any individual MT system output
- Exploit the complementarity of different MT approaches (rule-based, phrase-based, hierarchical, and syntax-based, etc...)
  $\Rightarrow$to produce consensus translations in the hope of generating better translations.

Existing Works

- Hypothesis selection using nbest list reranking based on various features [Hildebrand and Vogel, AMTA'08]
- Syscomb with SMT system, by considering source text and system outputs as bitext [Chen et al., WMT'09]
- Confusion network decoding :
  - does not require deep n-best lists
  - operates on the surface strings
  - [Rosti et al., ACL'07] [Shen et al., IWSLT'08] [Karakos et al., HLT'08] and [Matusov et al., EACL'06]

Architecture of MANY

Confusion Network (CN) based MT system combination



- Alignment of 1-best hypotheses and construction of CNs
- Construction of a lattice by merging CNs
- Decoding of the lattice
  - LM probability, Word penalty, Null-arc penalty, System weights.

Limits

- Exponential number of hypotheses $\rightsquigarrow$ n-grams do not appear in the system outputs $\rightsquigarrow$ ungrammatical
- LM score $\rightsquigarrow$ insufficient to precisely evaluate the hypotheses

Limits

- Exponential number of hypotheses $\rightsquigarrow$ n-grams do not appear in the system outputs $\rightsquigarrow$ ungrammatical
- LM score $\rightsquigarrow$ insufficient to precisely evaluate the hypotheses

Solutions

- To boost n-grams which appear in the system outputs
- Word confidence score

Plan

Two approaches to boost *n*-grams

- *n*-gram count feature : number of *n*-grams present in input hypotheses for each combined hypothesis
  - *bi* and *tri*-grams
- Adapted language model to decode the lattice
  - enhance the training data of a language model by the systems outputs
  - modify certain n-grams probabilities

Plan

Related work

- [Quirk.2004] : presents a supervised method for training a sentence level confidence measure on translation output using a human annotated corpus
- [Ueffing and Ney.2007] : present confidence scores at word-level based on word posterior probabilities
- [Hildebrand.2008] : defines several features extracted from *n*-best lists (at the sentence level) to select the best hypothesis in a combination approach via hypothesis selection.

Related work

- [Quirk.2004] : presents a supervised method for training a sentence level confidence measure on translation output using a human annotated corpus
- [Ueffing and Ney.2007] : present confidence scores at word-level based on word posterior probabilities
- [Hildebrand.2008] : defines several features extracted from *n*-best lists (at the sentence level) to select the best hypothesis in a combination approach via hypothesis selection.

$\Rightarrow$ Exploit certain features defined by *Hildebrand* to estimate a confidence score at the *word level* and injecting it into the confusion networks.

Confidence scores

1. Word agreement score based on a window of size **t** around position **i** ($WA_k(e_{i,t})$)
   - The relative frequency
2. Position independent *n*-best List *n*-gram Agreement ($NA_k(e_i)$) :
   - The percentage
3. N-best list *n*-gram probability ($NP_k(e_i)$):
   - *n*-gram language model probability

Confidence scores

1. Word agreement score based on a window of size **t** around position **i** ($WA_k(e_{i,t})$)
   - The relative frequency
2. Position independent n-best List n-gram Agreement ($NA_k(e_i)$) :
   - The percentage
3. N-best list n-gram probability ($NP_k(e_i)$):
   - n-gram language model probability

Word confidence score

$$SC_k(e_i) = \frac{\mathrm{WA}_k(e_i) + \sum_{j \in NG} \mathrm{NA}_k(e_i)^j + \mathrm{NP}_k(e_i)^j}{1 + 2 * |NG|} \quad (1)$$

$\Rightarrow$ NG={2-gram, 3-gram}

$\Rightarrow$ t = 2.

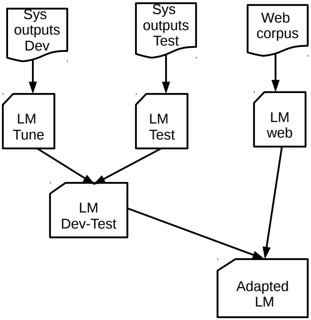
Plan

Data description

- The BOLT project on the Chinese to English translation task
- Outputs of six systems
  - 200-best lists $\Longrightarrow$ word confidence score
  - 1-best outputs $\Longrightarrow$ combination
- Corpora :

NAME	#sent.	#words.
Syscomtune	985	28671
Dev	1124	26350

Adapted language model



Language model	perplexity
LM-Web	295.43
Adapted-LM	169.923

Tests

- New features tests
  1. *n*-gram count
  2. Word confidence score
  3. *n*-gram count + Word confidence score

Tests

- New features tests
  1. *n*-gram count $\Longrightarrow$ boost-ngram
  2. Word confidence score $\Longrightarrow$ CS-ngram
  3. *n*-gram count + Word confidence score $\Longrightarrow$ Boost-ngram+CS-ngram

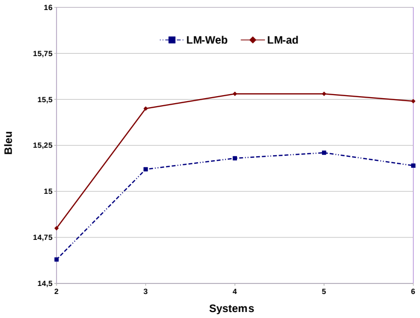
Tests

- New features tests
  1. *n*-gram count $\Longrightarrow$ boost-ngram
  2. Word confidence score $\Longrightarrow$ CS-ngram
  3. *n*-gram count + Word confidence score $\Longrightarrow$ Boost-ngram+CS-ngram

- The baseline combination system and each test are evaluated
  - *LM-Web*
  - *LM-ad*

Evaluation results

- LM-Web Vs LM-ad :

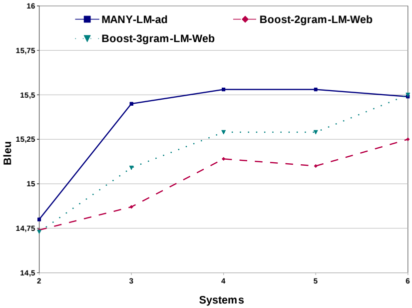


Systems	%BLEU
Sys1(Best)	**14.36**
Sys2	14.21
Sys3	13.76
Sys4	13.52
Sys5	13.36
Sys6	12.99
MANY+LM-Web	15.21
MANY+LM-ad	15.53

⇒ *0.85* and *1.17* %BLEU point relatively to the best single system
⇒ *MANY-LM-Web* is the baseline

Evaluation results

- Two approaches to boost n-grams :



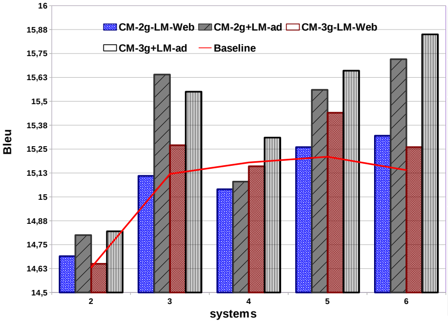
Systems	%BLEU
Sys1(Best)	14.36
MANY+LM-Web(baseline)	15.14
Boost-2gram+LM-Web	15.25
Boost-3gram+LM-Web	15.50
MANY+LM-ad	15.49
Boost-2gram+LM-ad	15.24
Boost-3gram+LM-ad	15.32

⇒ The adapted LM is better than *n*-gram count features to boost *n*-grams

⇒ LM-ad+n-gram count feature decrease results

Evaluation results

- Word confidence score :



Systems	%BLEU
Sys1(Best)	14.36
baseline	15.14
CS-2gram+LM-Web	15.25
CS-3gram+LM-Web	15.32
MANY+LM-ad	15.49
CS-2gram+LM-ad	15.72
CS-3gram+LM-ad	**15.85**

⇒ Contributes the most to the improvement of results
⇒ Performs better with the adapted LM than *LM-Web*
⇒ *1.49* and *0.71* %BLEU point over the best single system and the baseline

Evaluation results

- *n*-gram-count + word confidence score :

Systems	%BLEU
Sys1(Best)	14.36
Baseline	15.14
CS-2gram+LM-Web	15.25
CS-3gram+LM-Web	15.32
Boost+CS(2g)+LM-Web	**15.39**
Boost+CS(3g)+LM-Web	**15.78**
MANY+LM-ad	15.49
CS-2gram+LM-ad	15.72
CS-3gram+LM-ad	15.85
Boost+CS(2g)+LM-ad	**15.61**
Boost+CS(3g)+LM-ad	**15.74**

Conclusions

- MANY was run on six MT systems of different types
- An adapted LM and new features (n-gram count and confidence score) gave significant gains
- The use of an *adapted* LM in rescoring with word confidence score and the previews features improves results in term of BLEU score.
- The use of the two approaches to boost *n*-grams (*n*-gram count features and the adapted language model) together decreases results, this is mainly due to the redundancy

Perspectives

- Combine K-best hypotheses
  - complicate the search space
    $\Longrightarrow$ Reduce the number of backbones
    $\Longrightarrow$ The MBR method [Rosti et al., 2007].