Local System Voting Feature for Machine Translation System Combination

Markus Freitag, Jan-Thorsten Peter, Stephan Peitz, Minwei Feng and Hermann Ney

17. September 2015

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany
1 System Combination

▶ combine the output of multiple strong systems to one hypothesis

▶ combination confusion network approach (used by e.g. BBN, IBM, JHU)
  ▶ combine confusion networks built from the individual system outputs
  ▶ confusion network scored by several models
  ▶ decoding similar phrase-based machine translation decoders

▶ Successfully applied in several evaluation campaigns
  e.g. WMT [Freitag & Peitz⁺ 14], IWSLT [Freitag & Peitz⁺ 13],
  NTCIR [Feng & Freitag⁺ 13], WMT [Peitz & Mansour⁺ 13], WMT [Freitag & Peitz⁺ 12]

▶ Part of open source statistical machine translation toolkit Jane
Confusion Network Generation

- Select one of the input hypotheses as primary hypothesis
- Primary hypothesis determines the word order
  - All remaining hypotheses are word-to-word aligned
- Pairwise alignments generated via GIZA++
- The confusion network can be constructed with the calculated alignment
Decoding

- Do not stick to one primary hypothesis

- Final network is a union of all \( m \) (= amount individual systems) confusion networks (each having a different system as primary system)

- Final Network is scored by \( M \) models in a log-linear framework

  \[
  \sum_{i=1}^{M} \lambda_i h_i
  \]

- Scaling factors optimized with MERT on \( n \)-best lists

- Shortest path algorithm to extract final hypothesis

- All graph operations are conducted with openFST [Allauzen & Riley+ 07]
Features

- $m$ binary system voting features
  - For each word the voting feature for system $i$ ($1 \leq i \leq m$) is 1 iff the word is from system $i$, otherwise 0

- Binary primary system feature
  - Feature that marks the primary hypothesis

- LM feature
  - 3-gram language model trained on the input hypotheses

- Word penalty
  - Counts the number of words
2 Local System Voting Feature

Motivation:
- Binary voting features give preference to one or few individual systems
- Hypotheses with low voting feature weights have no effect on the final output

Idea:
- Define a local voting feature which give a score based on the current sentence/words
- Train model by a feed-forward neural network (NN) to give also unseen events a reliable score
- Related work from speech recognition: [Hillard & Hoffmeister 07] trained a classifier to learn which word should be selected
Neural Network Unigram Input Example

Best S BLEU path is labeled red

1-of-\( n \) encoding was applied to map words to a suitable NN input
Neural Network Bigram Input Example

Taking history of the individual hypotheses into account

1-of-\(n\) encoding was applied to map words to a suitable NN input
Neural Networks in System Combination

▶ Add one additional model based to the log-linear framework

▶ Training data:
  ▶ Split tuning set into 2 sets (one for NN training, one for MERT)
  ▶ Training samples cover only limited vocabulary
    ⇒ Use word classes

▶ Trainied using NPLM [Vaswani & Zhao 13]
# BOLT Arabic → English Results

| system combination                        | word classes | tune | test |
|-------------------------------------------|--------------|------|------|
| baseline                                  |              |      |      |
|                                           |              | 30.1 | 51.2 | 27.6 | 55.8 |
| +unigram neural network model             | no           |      |      |
|                                           | yes          | 31.4 | 51.2 | 28.5 | 56.0 |
|                                           |              | 31.1 | 51.1 | 28.3 | 55.7 |
| +bigram neural network model              | no           |      |      |
|                                           | yes          | 31.3 | 51.1 | 28.4 | 55.8 |
|                                           |              | 31.4 | 51.2 | 28.7 | 56.0 |

- 5 Systems
- 1510 sentences result in 6.5M training samples
- Test set has a OOV rate of 43.25%
- MERT tune set has a OOV rate of 43.24%
### BOLT Chinese→English Results

| system combination                | word classes | tune      | test       |
|-----------------------------------|--------------|-----------|------------|
|                                   |              | BLEU  | TER | BLEU  | TER |
| baseline                          |              | 17.9  | 61.5 | 18.3  | 60.9 |
| +unigram neural network model     | no           | 18.1  | 61.2 | 18.3  | 60.3 |
|                                   | yes          | 18.4  | 61.5 | 19.0  | 60.3 |
| +bigram neural network model      | no           | 18.1  | 61.3 | 18.6  | 60.3 |
|                                   | yes          | 18.1  | 61.2 | 18.7  | 59.9 |

- 9 Systems
- 1844 sentences result in 15M training samples
- Test set has a OOV rate of 40.73%
- MERT tune set has a OOV rate of 40.91%
BOLT Chinese→English Analysis

| #   | baseline     | +bigram wcNN  |
|-----|--------------|--------------|
| 1   | 120/14072    | 214/14072    |
|     | (0.9%)       | (1.5%)       |
| 2   | 592/6129     | 764/6129     |
|     | (9.7%)       | (12.5%)      |
| 3   | 1141/4159    | 1319/4159    |
|     | (27.4%)      | (31.7%)      |
| 4   | 1573/3241    | 1669/3241    |
|     | (48.5%)      | (51.5%)      |
| 5   | 2051/2881    | 1993/2881    |
|     | (71.2%)      | (69.2%)      |
| 6   | 2381/2744    | 2332/2744    |
|     | (86.8%)      | (85.0%)      |
| 7   | 2817/2965    | 2820/2965    |
|     | (95.0%)      | (95.1%)      |
| 8   | 3818/3860    | 3815/3860    |
|     | (98.9%)      | (98.8%)      |
| 9   | 11008/11008  | 11008/11008  |
|     | (100.0%)     | (100.0%)     |

More words created by a single or a few systems are used
3 Conclusion

- Proposed novel local system voting model
- Using feedforward neural network models
- Allow confusion network to prefer other systems even in the same sentence
- Improved likelihood to select words created by only few systems
- Use word classes to avoid sparsity problem
- Improvements of 0.7% for Ch-En and 1.1% for Ar-En
Thank you for your attention

Markus Freitag, Jan-Thorsten Peter, Stephan Peitz, Minwei Feng and Hermann Ney

surname@cs.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/
References

[Allauzen & Riley+ 07] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, M. Mohri. OpenFst: A General and Efficient Weighted Finite-State Transducer Library. In J. Holub, J. Zdárek, editors, *Implementation and Application of Automata*, Vol. 4783 of *Lecture Notes in Computer Science*, pp. 11–23. Springer Berlin Heidelberg, 2007. 4

[Feng & Freitag+ 13] M. Feng, M. Freitag, H. Ney, B. Buschbeck, J. Senellart, J. Yang. The system combination rwth aachen: Systran for the ntcir-10 patentmtm evaluation. In *10th NTCIR Conference*, pp. 301–308, Tokyo, Japan, June 2013. 2

[Freitag & Peitz+ 12] M. Freitag, S. Peitz, M. Huck, H. Ney, T. Herrmann, J. Niehues, A. Waibel, A. Allauzen, G. Adda, B. Buschbeck, J. M. Crego, J. Senellart. Joint wmt 2012 submission of the quaero project. In *NAACL 2012 Seventh Workshop on Statistical Machine Translation (WMT)*, pp. 322–329, Montreal, Canada, June 2012. 2

[Freitag & Peitz+ 13] M. Freitag, S. Peitz, J. Wuebker, H. Ney, N. Durrani, M. Huck, P. Koehn, T.-L. Ha, J. Niehues, M. Mediani, T. Herrmann, A. Waibel, N. Bertoldi, M. Cettolo, M. Federico. Eu-bridge mt: Text translation of talks in the eu-bridge
project. In *International Workshop on Spoken Language Translation (IWSLT)*, pp. 128–135, Heidelberg, Germany, December 2013. 2

[Freitag & Peitz + 14] M. Freitag, S. Peitz, J. Wuebker, H. Ney, M. Huck, R. Sennrich, N. Durrani, M. Nadejde, P. Williams, P. Koehn, T. Herrmann, E. Cho, A. Waibel. Eu-bridge mt: Combined machine translation. In *ACL 2014 Ninth Workshop on Statistical Machine Translation (WMT)*, pp. 105–113, Baltimore, Maryland, USA, June 2014. 2

[Hillard & Hoffmeister + 07] D. Hillard, B. Hoffmeister, M. Ostendorf, R. Schlüter, H. Ney. i rover: improving system combination with classification. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 65–68, Rochester, NY, USA, April 2007. Association for Computational Linguistics. 6

[Peitz & Mansour + 13] S. Peitz, S. Mansour, M. Huck, M. Freitag, H. Ney, E. Cho, T. Herrmann, M. Mediani, J. Niehues, A. Waibel, A. Allauzen, Q. K. Do, B. Buschbeck, T. Wandmacher. Joint wmt 2013 submission of the quaero project. In *Eighth Workshop on Statistical Machine Translation (WMT)*, pp. 185–192, Sofia, Bulgaria, August 2013. 2

[Vaswani & Zhao + 13] A. Vaswani, Y. Zhao, V. Fossum, D. Chiang. Decoding with large-scale neural language models improves translation. In *Conference on
BOLT Arabic→English System

|                     | Arabic | English |
|---------------------|--------|---------|
| Sentences           |        | 8M      |
| Running words       | 189M   | 186M    |
| Vocabulary          | 608K   | 519K    |
| Tune sentences      | 1510 (NN), 1080 (MERT) |         |
| Test sentences      |        | 1137    |

5 Systems
1510 sentences result in 6.5M training samples
Test set has a OOV rate of 43.25% MERT tune set has a OOV rate of 43.24%
BOLT Chinese→English Systems

|                          | Chinese | English |
|--------------------------|---------|---------|
| Sentences                | 13M     |         |
| Running words            | 255M    | 279M    |
| Vocabulary               | 370K    | 833K    |
| Tune sentences           | 1844 (NN), 985 (MERT) |         |
| Test sentences           |         | 1124    |

9 Systems

1844 sentences result in 15M training samples
Test set has a OOV rate of 40.73% MERT tune set has a OOV rate of 40.91%