CMU-StatXfer Group System Combination

Kenneth Heafield

Language Technologies Institute
Carnegie Mellon University

September 1, 2009

Submissions

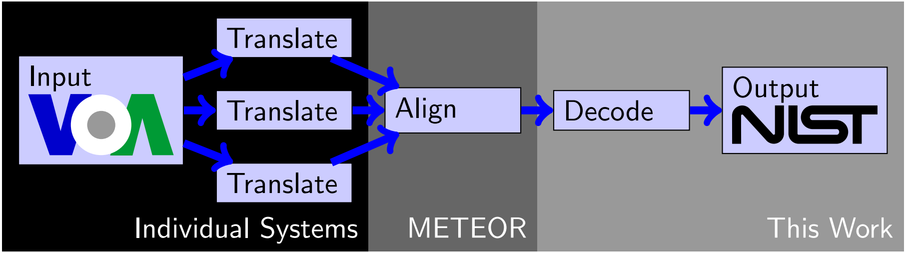
Formal System Combination

- Urdu-English using:
  - AFRL
  - JHU: Joshua decoder
  - CMU-StatXfer primary: Moses decoder
  - CMU-StatXfer contrast2: Xfer decoder

Informal System Combination

- Arabic-English
- Urdu-English

Pipeline


Input
Translate
Translate
Translate
Align
Decode
Output
Individual Systems
METEOR
This Work

Arabic-English Example Combination

System 1: So even if that was meaningful , it is because you were late

System 2: Even if feasible , it is because you have been delayed

$\Downarrow$ Combine

Combined: Even if feasible , it is because you were late

$\neq$ Compare

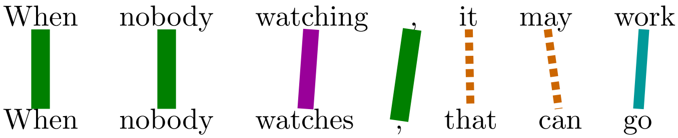
Reference: And even if that was useful , it was because you were late

Outline

1. Alignment
2. Search Space
3. Features
   - Support
   - Tuning

Sentence Pair Alignment

Match surface, stems, and WordNet synsets
Minimize crossing alignments
Speculate using part of speech when neighbors align



Lavie and Agarwal, METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments, WMT 2007.

Overall Alignment: Urdu-English Example

1 Russian President Putin Mir ولادی it for a big success .

2 The Russian president ولادی the result of a big victory for Putin .

Overall Alignment: Urdu-English Example

1 Russian President Putin Mir ولادی it for a big success .

2 The Russian president ولادی the result of a big victory for Putin .

1 Russian President Putin Mir ولادی it for a big success .

3 For the result Russian President ولادی Mir Putin is a great success .

2 The Russian president ولادی the result of a big victory for Putin .

3 For the result Russian President ولادی Mir Putin is a great success .

Alignment Comparison with Confusion Networks

	Confusion Networks	**This Work**
Alignment Method	TER or ITG	METEOR
Sentences Aligned	To Skeleton(s)	All Pairs

Outline

1. Alignment
2. Search Space
3. Features
   - Support
   - Tuning

Search Space

Algorithm

Start at the beginning of each sentence

Branch by appending the first unused word from a system

Example

System 1: Now can know why .

System 2: Now we can now know why .

Partial Hypothesis

$$\begin{cases} \text{Now} \\ \text{Now} \end{cases}$$

Search Space

Algorithm

Start at the beginning of each sentence

Branch by appending the first unused word from a system

Use the appended word and those aligned with it

Example

System 1: Now can know why .

System 2: Now we can now know why .

Partial Hypothesis

Now {can, we}

Search Space

Algorithm

Start at the beginning of each sentence
Branch by appending the first unused word from a system
Use the appended word and those aligned with it
Loop until all hypotheses reach end of sentence

Example

System 1: Now can know why .
System 2: Now we can now know why .

Partial Hypothesis

Now we $\begin{cases} \text{can} \\ \text{can} \end{cases}$

Search Space

Algorithm

Start at the beginning of each sentence

Branch by appending the first unused word from a system

Use the appended word and those aligned with it

Loop until all hypotheses reach end of sentence

Example

System 1: Now can know why .

System 2: Now we can now know why .

Partial Hypothesis

Now we can $\begin{cases} \text{know} \\ \text{now} \end{cases}$

Search Space Comparison with Confusion Networks

	Confusion Networks	**This Work**
Inputs	*n*-best	1-best
Word Ordering	Skeleton	Switches Every Word

One Interpretation

Confusion network that dynamically switches skeletons

Outline

1. Alignment
2. Search Space
3. Features
    - Support
    - Tuning

Features

Length

Length of hypothesis

Language Model

Model: log probability from SRI language model

***n*-Gram:** length of *n*-gram found in model

Support

Count of *n*-grams supported by each system

Support Features

System 1: Supported Proposal of France

System 2: Support for the Proposal of France

↓ Hypothesis

Hypothesis: Support for Proposal of France

↓ Count

	Unigram	**Bigram**	**Trigram**	**Quadgram**
System 1	4	2	1	0
System 2	5	3	1	0

Rationale for Support Features

Confidence

Tuned feature weights are confidence in each system.

Language Model On Inputs

Simple language model trained on inputs and tuned using MERT.

Impact on BLEU

Systems vote on n-grams which BLEU evaluates.

Comparison of Support Features

System Weights	**Sites**
Uniform	Hildebrand, IBM, JHU, TUBITAK
Rank	SRI, Zhao
BLEU	BBN, HIT-LTRC
Tuned	BBN, RWTH, Zens, This Work

Comparison of Support Features

System Weights	Sites
Uniform	Hildebrand, IBM, JHU, TUBITAK
Rank	SRI, Zhao
BLEU	BBN, HIT-LTRC
Tuned	BBN, RWTH, Zens, This Work

n-**Gram Weights**	**Sites**
Unigram Only	BBN, HIT-LTRC, SRI
Constant	IBM, JHU, RWTH, TUBITAK, Zens, Zhao
Tuned	Hildebrand, This Work

Parameter Tuning

Overall Score

Linear combination of length, language model, and support features

Tuning

Minimum Error Rate Training for feature weights

Too Many Features

Arabic Numbers

Systems Combined	9
Features	39
Tuning Segments	317

Problems

- MERT instability
- Overfitting

Reduce the Features

System Weights

Tuned system weights for short *n*-grams
Uniform system weights for long *n*-grams

Features	Uncased Tune	Cased Tune	Cased Test	Submission
15	57.65	55.68	53.75	contrast2
23	59.50	57.60	55.30	primary
39	58.88	56.92	55.12	contrast1

Table: Arabic BLEU scores by number of features

Reduce the Features

Tuning BLEU decreased by 0.62 with more features.

System Weights

Tuned system weights for short *n*-grams
Uniform system weights for long *n*-grams

Features	Uncased Tune	Cased Tune	Cased Test	Submission
15	57.65	55.68	53.75	contrast2
23	59.50	57.60	55.30	primary
39	58.88	56.92	55.12	contrast1

Table: Arabic BLEU scores by number of features

Tuned System Weights

Best system has highest weight.

System	BLEU	Unigram	Bigram
17	51.72	4.3669	16.5329
08	51.49	0.8562	2.5201
14	50.28	2.5157	0.0197
06	49.42	0.3316	6.5232
16	49.38	0.6493	0.3347
02	49.30	0.9713	2.5741
07	49.15	0.2788	0.8149
03	47.90	2.2679	1.5260
01	47.43	0.5319	1.3003

Table: Tuned unigram and bigram weights for Arabic primary submission. BLEU is uncased on the system combination tuning set.

Tuned System Weights

Weight is not monotonic by BLEU.

System	BLEU	Unigram	Bigram
17	51.72	4.3669	16.5329
08	51.49	0.8562	2.5201
14	50.28	2.5157	0.0197
06	49.42	0.3316	6.5232
16	49.38	0.6493	0.3347
02	49.30	0.9713	2.5741
07	49.15	0.2788	0.8149
03	47.90	2.2679	1.5260
01	47.43	0.5319	1.3003

Table: Tuned unigram and bigram weights for Arabic primary submission. BLEU is uncased on the system combination tuning set.

Tuned System Weights

Individual trade-off between unigrams and bigrams.

System	BLEU	Unigram	Bigram
17	51.72	4.3669	16.5329
08	51.49	0.8562	2.5201
14	50.28	2.5157	0.0197
06	49.42	0.3316	6.5232
16	49.38	0.6493	0.3347
02	49.30	0.9713	2.5741
07	49.15	0.2788	0.8149
03	47.90	2.2679	1.5260
01	47.43	0.5319	1.3003

Table: Tuned unigram and bigram weights for Arabic primary submission. BLEU is uncased on the system combination tuning set.

Hyperparameter Tuning

Hyperparameters

- Set of systems combined
- Number of support features
- Synchronization method

Brute Force

Decoder does 2.9 combinations/second, so I tried and fully tuned 63 combinations.

Timed on a Core 2 Quad 2.83GHz with 9 Arabic systems

Outline

4 Conclusion
- Results
- References and Acknowledgments

Formal Urdu-English

Urdu-English

1.82 BLEU gain

BLEU	Submission
25.04	Combined primary

BLEU	Component Systems
23.22	CMU-StatXfer primary: Moses decoder
22.93	JHU Joshua
22.35	AFRL
16.00	CMU-StatXfer contrast2: Xfer decoder

Case-sensitive BLEU

Informal Results

Urdu-English

1.24 BLEU gain

BLEU	Submission
32.28	contrast3
31.88	primary
31.71	contrast1
31.62	contrast2

BLEU	Best Component
31.04	System 9

Arabic-English

5.22 BLEU gain

BLEU	Submission
55.30	primary
55.25	contrast3
55.12	contrast1
53.75	contrast2

BLEU	Best Component
50.08	System 8 unconstrained

Case-sensitive BLEU

References

- Hildebrand and Vogel, Combination of Machine Translation Systems via Hypothesis Selection from Combined N-Best Lists, AMTA 2008.
- Zens and Ney, N-Gram Posterior Probabilities for Statistical Machine Translation, WMT 2006.
- Zhao and He, Using N-gram based Features for Machine Translation System Combination, NAACL HLT 2009.

Acknowledgments

Greg Hanneman previous maintainer of the system
Alon Lavie adviser
Alok Parlikar language model for formal system combination
Mohamed Noamany language model for informal system combination