Results of the WMT16 Tuning Shared Task

Bushra Jawaid
Amir Kamran
Miloš Stojanović
Ondřej Bojar

1ILLC, University of Amsterdam
2MFF UFAL, Charles University in Prague

WMT16, Aug 11, 2016
Overview

• Summary of Tuning Task

• Updates in 2016 edition

• Results
Tuning Task

\[ \hat{t} = \arg\max_{t \in T(s)} \lambda \phi(s, t) \]
Tuning Task

\[ \hat{t} = \arg\max_{t \in T(s)} \lambda \phi(s, t) \]
Tuning Task

\[ \hat{t} = \arg \max_{t \in T(s)} \lambda \phi(s, t) \]
Tuning Task

\[ \hat{t} = \arg\max_{t \in T(s)} \lambda \phi(s, t) \]
Tuning Task

\[ \hat{t} = \arg \max_{t \in T(s)} \lambda \phi(s, t) \]

So many things to choose in tuning:
Tuning Task

\[
\hat{t} = \underset{t \in T(s)}{\text{argmax}} \lambda \phi(s, t)
\]

So many things to choose in tuning:

- Metric
- Algorithm
- Data
- Features
- ...
Tuning Task

\[
\hat{t} = \arg\max_{t \in T(s)} \lambda\phi(s, t)
\]

So many things to choose in tuning:

- Metric
- Algorithm
- Data
- Features
- ...

This task is organized to explore the tuning options in a controlled settings
System for Tuning

- Moses phrase-based models trained both for English-Czech and Czech-English.
System for Tuning

- Moses phrase-based models trained both for English-Czech and Czech-English.

- This year we used large dataset to train the models and aligned the data using fast-align.
System for Tuning

- Moses phrase-based models trained both for English-Czech and Czech-English.

- This year we used large dataset to train the models and aligned the data using fast-align.

- In constrained version 2.5K sentence pairs were available for tuning.
System for Tuning

• Moses phrase-based models trained both for English-Czech and Czech-English.

• This year we used large dataset to train the models and aligned the data using fast-align.

• In constrained version 2.5K sentence pairs were available for tuning.

• Constrained version allowed only dense features.
System for Tuning

- Moses phrase-based models trained both for English-Czech and Czech-English.
- This year we used large dataset to train the models and aligned the data using fast-align.
- In constrained version 2.5K sentence pairs were available for tuning.
- Constrained version allowed only dense features.
- Any tuning algorithm or metric tuning was allowed (even manually setting weights)
## Data used for training

| Source          | Sentences | Tokens | Types |
|-----------------|-----------|--------|-------|
| **LM Corpora**  |           |        |       |
| Europarl v7, News Commentary v11, News Crawl (2007-15), News Discussion v1 | 54M | 900M | 2.1M |
|                 | 206M      | 4409M  | 3.2M  |
| **TM Corpora**  |           |        |       |
| CzEng 1.6 pre for WMT16 | 44M | 501M | 1.8M |
|                 |           | 20.8M  | 1.2M  |
| **Dev Set**     |           |        |       |
| newstest2015    | 2656      | 51K    | 19K   |
|                 |           | 60K    | 13K   |
| **Test Set**    |           |        |       |
| newstest2016    | 2999      | 56.9K  | 15.1K |
|                 |           | 65.3K  | 8.8K  |
Data used for training

Comparison of data sizes (# of sentence pairs) 2015 vs 2016
Participants

- From 6 research groups we received, 4 submissions for Czech-English, 8 submissions for English-Czech
- 2 Baselines

| System                  | Participant                                                                 |
|-------------------------|-----------------------------------------------------------------------------|
| bleu-MIRA, bleu-MERT    | Baselines                                                                   |
| AFRL                    | United States Air Force Research Laboratory                                |
| DCU                     | Dublin City University                                                     |
| FJFI-PSO                | Czech Technical University in Prague                                        |
| ILLC-UvA-BEER           | ILLC – University of Amsterdam                                              |
| NRC-MEANT, NRC-NNBLEU   | National Research Council Canada                                            |
| USAAR                   | Saarland University                                                        |
## Czech-English Results

| System Name           | True Skill Score | BLEU  |
|-----------------------|------------------|-------|
| BLEU-MIRA             | 0.114            | 22.73 |
| AFRL                  | 0.095            | 22.90 |
| NRC-NNBLEU            | 0.090            | 23.10 |
| NRC-MEANT             | 0.073            | 22.60 |
| ILLC-UvA-BEER         | 0.032            | 22.46 |
| BLEU-MERT             | 0.000            | 22.51 |
## Czech-English Results

| System Name       | True Skill Score | BLEU  |
|-------------------|------------------|-------|
| BLEU-MIRA         | 0.114            | 22.73 |
| AFRL              | 0.095            | 22.90 |
| NRC-NNBLEU        | 0.090            | 23.10 |
| NRC-MEANT         | 0.073            | 22.60 |
| ILLC-UvA-BEER     | 0.032            | 22.46 |
| BLEU-MERT         | 0.000            | 22.51 |

- Manual evaluation of tuning systems can draw only very few clear division lines.
Czech-English Results

| System Name       | True Skill Score | BLEU  |
|-------------------|------------------|-------|
| BLEU-MIRA         | 0.114            | 22.73 |
| AFRL              | 0.095            | 22.90 |
| NRC-NNBLEU        | 0.090            | 23.10 |
| NRC-MEANT         | 0.073            | 22.60 |
| ILLC-UvA-BEER     | 0.032            | 22.46 |
| BLEU-MERT         | 0.000            | 22.51 |

- Manual evaluation of tuning systems can draw only very few clear division lines.
- KBMIRA turns out to consistently be better than MERT.
## English-Czech Results

| System Name                     | True Skill Score | BLEU  |
|---------------------------------|------------------|-------|
| BLEU-MIRA                       | 0.160            | 15.12 |
| ILLC-UvA-BEER                   | 0.152            | 14.69 |
| BLEU-MERT                       | 0.151            | 14.93 |
| AFRL2                           | 0.139            | 14.84 |
| AFRL1                           | 0.136            | 15.02 |
| DCU                             | 0.134            | 14.34 |
| FJFI-PSO                        | 0.127            | 14.68 |
| USAAR-HMM-MERT                  | -0.433           | 7.95  |
| USAAR-HMM-MIRA                  | -1.133           | 0.82  |
| USAAR-HMM                       | -1.327           | 0.20  |
Comparison with main translation task

| #  | score | range | system           |
|----|-------|-------|------------------|
| 1  | 0.62  | 1     | UEDIN-NMT       |
| 2  | 0.32  | 2     | JHU-PBMT        |
| 3  | 0.21  | 3     | ONLINE-B        |
| 4  | 0.11  | 4-6   | TT-BLEU-MIRA    |
|    | 0.10  | 4-7   | TT-AFRL         |
|    | 0.09  | 4-7   | TT-NRC-NNBLEU   |
|    | 0.07  | 5-8   | TT-NRC-MEANT    |
|    | 0.03  | 7-10  | TT-BEER-PRO     |
|    | 0.00  | 8-10  | PJATK            |
|    | 0.00  | 8-10  | TT-BLEU-MERT    |
| 5  | -0.07 | 11    | ONLINE-A        |
| 6  | -1.48 | 12    | CU-MRGTREES     |

| #  | score | range | system           |
|----|-------|-------|------------------|
| 1  | 0.619 | 1     | ONLINE-B        |
| 2  | 0.574 | 2     | UEDIN-JHU       |
| 3  | 0.532 | 3-4   | UEDIN-SYNTAX    |
|    | 0.518 | 3-4   | MONTREAL        |
| 4  | 0.436 | 5     | ONLINE-A        |
| 5  | -0.125| 6     | CU-TECTO        |
| 6  | -0.182| 7-9   | TT-BLEU-MIRA-D  |
|    | -0.189| 7-10  | TT-ILLC-UVA     |
|    | -0.196| 7-11  | TT-BLEU-MERT    |
|    | -0.210| 8-11  | TT-AFRL         |
|    | -0.220| 9-11  | TT-USAAR-TUNA   |
| 7  | -0.263| 12-13 | TT-DCU          |
|    | -0.297| 13-15 | TT-METEOR-CMU   |
|    | -0.320| 13-15 | TT-BLEU-MIRA-SP |
|    | -0.320| 13-15 | TT-HKUST-MEANT  |
|    | -0.358| 15-16 | ILLINOIS        |
## Comparison with main translation task

### English–Czech

| #  | score | range | system           |
|----|-------|-------|------------------|
| 1  | 0.59  | 1     | UEDIN-NMT        |
| 2  | 0.43  | 2     | NYU-MONTREAL     |
| 3  | 0.34  | 3     | JHU-PBMT         |
| 4  | 0.30  | 4-5   | CU-CHIMERA       |
|    | 0.30  | 4-5   | CU-TAMCHYNA      |
| 5  | 0.22  | 6-7   | UEDIN-CU-SYTX    |
|    | 0.19  | 6-7   | ONLINE-B         |
| 6  | 0.16  | 8-11  | TT-BLEU-MIRA     |
|    | 0.15  | 8-12  | TT-BEER-PRO      |
|    | 0.15  | 8-13  | TT-BLEU-MERT     |
|    | 0.14  | 9-14  | TT-AFRL2         |
|    | 0.14  | 9-14  | TT-AFRL1         |
|    | 0.13  | 9-14  | TT-DCU           |
|    | 0.13  | 11-14 | TT-FJFI          |
| 7  | 0.08  | 15    | ONLINE-A         |
| 8  | -0.03 | 16    | CU-TECTOTMT      |
| 9  | -0.43 | 17    | TT-USAAR-HMM-MERT|
| 10 | -0.54 | 18    | CU-MRGTREES      |
| 11 | -1.13 | 19    | TT-USAAR-HMM-MIRA|
| 12 | -1.33 | 20    | TT-USAAR-HARM    |

### English–Czech

| #  | score  | range | system           |
|----|--------|-------|------------------|
| 1  | 0.686  | 1     | CU-CHIMERA       |
| 2  | 0.515  | 2-3   | ONLINE-B         |
|    | 0.503  | 2-3   | UEDIN-JHU        |
| 3  | 0.467  | 4     | MONTREAL         |
| 4  | 0.426  | 5     | ONLINE-A         |
| 5  | 0.261  | 6     | UEDIN-SYNTAX     |
| 6  | 0.209  | 7     | CU-TECTO         |
| 7  | 0.114  | 8     | COMMERCIAL1      |
| 8  | -0.342 | 9-11  | TT-DCU           |
|    | -0.342 | 9-11  | TT-AFRL          |
|    | -0.346 | 9-11  | TT-BLEU-MIRA-D   |
| 9  | -0.373 | 12    | TT-USAAR-TUNA    |
| 10 | -0.406 | 13    | TT-BLEU-MERT     |
| 11 | -0.563 | 14    | TT-METEOR-CMU    |
| 12 | -0.808 | 15    | TT-BLEU-MIRA-SP  |
Conclusion

• The task was much larger this year.

• Task attracted good participation like last year.

• The quality of most submitted systems is hard to distinguish manually.

• With large models, the few parameters are most likely not powerful enough (and sadly nobody tried discriminative features).

• The results confirm that KBMIRA with the standard features optimized towards BLEU should be preferred over MERT.