Improved Statistical Machine Trasnlation by Harmonizing Source and Target Text Before Training and Testing

The purpose of this document is to illustrate how to replicate the experiment that I have made and which I was able to improve the BLEU score of a Arabic-English baseline SMT from 0.3104 to 0.3321.

Environment & Data

The experiments described here were done on the following environment: * Ubuntu 12.04 (64-bit) - We used EC2 instance of type m3.xlarge * GIZA 1.0.7 * SRILM 1.5.12 * Moses Decoder - Pulled latest from Github on 4/9/2013 * MADA 3.2 (with SVMTools 1.3.1, aramorph-1.2.1, SRLIM 1.5.12 not patched) * Python 2.7 * Additional Python Modules: numpy, scikits-learn

Data

- Training Data: 34k Sentence Pairs (Cleaned) - Roughly 996458 words according to MADA output > https://dl.dropbox.com/s/agv9khd6mvfa8as/TrainData.zip

- LM Model: 106930 Sentences > https://dl.dropbox.com/s/4wyaebj0dln2xgt/LM*data%2BTrain*data.en.zip

Baseline SMT Training & Evaluation

Training

Starting from project root directory, run the following:

```
# Start at project root dir
cd SMT/Baseline
mkdir -p work/LM
cp ../../LM_data+Train_data.en.lm work/LM/
$SCRIPTS_ROOTDIR/training/train-model.perl  -external-bin-dir /home/ubuntu/tools/bin \
                                            -root-dir work \
                                            -corpus data/Train/Train_data.clean \
                                            -f ar -e en -alignment grow-diag-final-and \
                                            -reordering msd-bidirectional-fe \
                                            -lm 0:3:/home/ubuntu/workspace/mt-arabic-english-har

mkdir -p work/tuning
$SCRIPTS_ROOTDIR/training/mert-moses.pl data/Tune/Tune_data.mt04.50.ar data/Tune/Tune_data.mt04.
$SCRIPTS_ROOTDIR/scripts/reuse-weights.perl work/tuning/mert/moses.ini < work/model/moses.ini >
```

Notes:

1. /home/ubuntu/tools/bin points to GIZA++ binaries directory
2. LM model must be absolute path

Evaluation:

To evaluate the SMT, run the following:

```
$SCRIPTS_ROOTDIR/training/filter-model-given-input.pl work/evaluation/filtered work/tuning/moses
/home/ubuntu/tools/moses/bin/moses -config work/evaluation/filtered/moses.ini -input-file data/T
$SCRIPTS_ROOTDIR/wrap-xml.perl data/Test/Test_data.mt05.ref.ar.xml en my-system-name < work/eval
$SCRIPTS_ROOTDIR/mteval-v11b.pl -s data/Test/Test_data.mt05.src.ar.xml -r data/Test/Test_data.mt
```

Last command should output the BLEU and NIST scores for the Baseline. In my case the output was (BLEU: 0.3104, NIST: 6.5153)

```
  Evaluation of Arabic-to-English translation using:
    src set "mt05 arabic evlset v0" (4 docs, 48 segs)
    ref set "mt05_arabic_evlset_v0-ref" (4 refs)
    tst set "mt05_arabic_evlset_v0" (1 systems)

NIST score = 6.5153  BLEU score = 0.3104 for system "ahd"

# ------------------------------------------------------------------------

Individual N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  5.1632   1.0553   0.2081   0.0583   0.0305   0.0077   0.0032   0.0018   0.0006  "ahd"

 BLEU:  0.7229   0.3936   0.2421   0.1484   0.0836   0.0480   0.0270   0.0155   0.0122  "ahd"

# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  5.1632   6.2185   6.4265   6.4848   6.5153   6.5230   6.5262   6.5280   6.5286  "ahd"

 BLEU:  0.7056   0.5207   0.4001   0.3104   0.2376   0.1813   0.1376   0.1045   0.0821  "ahd"
MT evaluation scorer ended on 2013 Apr 11 at 00:25:04
```

Improving the Baseline using Harmonizer

To improve the baseline, we would like to train a harmonizer that is capable of making Arabic source text more similar to the English target text by using single tokens for various Arabic surface forms that map to single English word. To achieve this, we (1) Train a Harmonizer and then (2) Use the Harmonizer to harmonize trainig parallal corpus, tuning data and test data. Note that once a harmonizer is trained, it needs not to be retrianed everytime to be reused.

1) Training the Harmonizer

To train the Harmonizer, we do the following:

1. Analyze the Arabic side of the training corpus and annotate it with lemmas and morphological features
2. Train a factored phrase table that maps lemmas+features to English words
3. Extract features where same lemma maps to same English form (Label "collapsible")
4. Extract features where same lemma maps to different English forms (Label "non-collapsible")
5. Train a binary classifier data from step 3 and step 4

The following sequence of commands perform the steps above, we are using the same parallal corpus and English LM to train our harmonizer, but a different data set can be also used to train the harmonizer:

```
# Start at project root dir
cd harmonizer

## Analyze Arabic side of text
# Make sure conf/template.madaconfig points to correct MADA installation directory
nohup perl $MADAHOME/MADA+TOKAN.pl config=conf/template.madaconfig file=data/Train/Train_data.cl

## Create an annotated corpus
python factorize-corpus.py data/Train/Train_data.clean.ar.bw.mada > data/Train/Train_data.clean.
cp data/Train/Train_data.clean.en data/Train/Train_data.clean.factored.en

mkdir -p work/LM
cp ../LM_data+Train_data.en.lm work/LM/

$SCRIPTS_ROOTDIR/training/train-model.perl  -external-bin-dir /home/ubuntu/tools/bin \
                                            -root-dir work \
                                            -corpus data/Train/Train_data.clean.factored \
                                            -f ar -e en -alignment grow-diag-final-and \
                                            -reordering msd-bidirectional-fe \
                                            -lm 0:3:/home/ubuntu/workspace/mt-arabic-english-har
                                            -alignment-factors 1-0 \
                                            -translation-factors 1,2-0

mkdir data/Harmonizer
python cluster-annotated-table.py work/model/phrase-table.1,2-0.gz > data/Harmonizer/harmonizer_
python train_harmonizer.py data/Harmonizer/harmonizer_training_data.csv true
```

Note: In my case, the annotated corpus yieleded 34634 total entries. 5322 labeled collapsible and 29312 non-collapsible (almost a ratio of 1 positive to 5 negative)

2) Using the Harmonizer to improve our data

Now that we have a Harmonizer ready to be used. We will use it to harmonize our training data and tuning data to build an improved SMT. **Note:** the harmonizer takes an annotated corpus text as input and produces the harmonized data. We will use the same annotated corpus text we used when trained our harmonizer to save time

```
# Start at project root dir

# Copy the annotated corpus we created when we trained the harmonizer
cp harmonizer/data/Train/Train_data.clean.factored.ar harmonizer/data/Train/Train_data.clean.fac

# Use the harmonizer to create a harmonized corpus from the annotated one
python harmonizer/harmonizer.py harmonizer/harmonizer_model.pkl SMT/Improved/data/Train/Train_da
cp SMT/Improved/data/Train/Train data.clean.en SMT/Improved/data/Train/Train data.clean.harmoniz
```

Now we have a harmonized parallel corpus in the following paths: > SMT/Improved/data/Train/Train*data.clean.harmonized.ar >*
SMT/Improved/data/Train/Train data.clean.harmonized.en

Let's repeat the same steps to harmonize our tuning data:

```
Start at project root dir

# Harmonize tuning data
perl $MADAHOME/MADA+TOKAN.pl config=harmonizer/conf/template.madaconfig file=SMT/Improved/data/T

# Create an annotated tuning data
python harmonizer/factorize-corpus.py SMT/Improved/data/Tune/Tune_data.mt04.50.ar.bw.mada > SMT/

# Use harmonizer on tuning data
python harmonizer/harmonizer.py harmonizer/harmonizer_model.pkl SMT/Improved/data/Tune/Tune_data
cp SMT/Improved/data/Tune/Tune data.mt04.50.en SMT/Improved/data/Tune/Tune data.mt04.50.harmoniz
```

And again for the test data:

```
Start at project root dir

# Harmonize test data
perl $MADAHOME/MADA+TOKAN.pl config=harmonizer/conf/template.madaconfig file=SMT/Improved/data/T

python harmonizer/factorize-corpus.py SMT/Improved/data/Test/Test_data.mt05.src.ar.bw.mada  > SM
python harmonizer/harmonizer.py harmonizer/harmonizer model.pkl SMT/Improved/data/Test/Test data
```

Building Improved SMT & Evaluation:

Now we have improved data, we can use it to build and evaluate our SMT the same we way we did for our Baseline

```
cd SMT/Improved
mkdir -p work/LM
cp ../../LM_data+Train_data.en.lm work/LM/
nohup $SCRIPTS_ROOTDIR/training/train-model.perl -external-bin-dir /home/ubuntu/tools/bin \
                                                -root-dir work \
                                                -corpus data/Train/Train_data.clean.harmonized \
                                                -f ar -e en -alignment grow-diag-final-and \
                                                -reordering msd-bidirectional-fe \
                                                -lm 0:3:/home/ubuntu/workspace/mt-arabic-english-har
```

```
# Tune our SMT
cd SMT/Improved
mkdir -p work/tuning
nohup $SCRIPTS_ROOTDIR/training/mert-moses.pl data/Tune/Tune_data.mt04.50.harmonized.ar data/Tun
$SCRIPTS_ROOTDIR/scripts/reuse-weights.perl work/tuning/mert/moses.ini < work/model/moses.ini >
```

Evaluation:

```
# Evaluate
cd SMT/Improved
$SCRIPTS_ROOTDIR/training/filter-model-given-input.pl work/evaluation/filtered work/tuning/moses
/home/ubuntu/tools/moses/bin/moses -config work/evaluation/filtered/moses.ini -input-file data/T
$SCRIPTS_ROOTDIR/wrap-xml.perl data/Test/Test_data.mt05.ref.ar.xml en my-system-name < work/eval
$SCRIPTS_ROOTDIR/mteval-v11b.pl -s data/Test/Test_data.mt05.src.ar.xml -r data/Test/Test_data.mt

MT evaluation scorer began on 2013 Apr 11 at 19:56:13
command line:  /home/ubuntu/tools/moses/scripts/mteval-v11b.pl -s data/Test/Test_data.mt05.src.a
  Evaluation of Arabic-to-English translation using:
    src set "mt05_arabic_evlset_v0" (4 docs, 48 segs)
    ref set "mt05_arabic_evlset_v0-ref" (4 refs)
    tst set "mt05_arabic_evlset_v0" (1 systems)

NIST score = 6.5534  BLEU score = 0.3321 for system "ahd"

# ------------------------------------------------------------------------

Individual N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  5.1455   1.0917   0.2286   0.0629   0.0247   0.0109   0.0071   0.0036   0.0031  "ahd"

 BLEU:  0.7318   0.4258   0.2744   0.1713   0.1023   0.0620   0.0381   0.0219   0.0157  "ahd"

# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  5.1455   6.2373   6.4658   6.5287   6.5534   6.5643   6.5714   6.5750   6.5781  "ahd"

 BLEU:  0.6986   0.5329   0.4206   0.3321   0.2600   0.2031   0.1589   0.1233   0.0976  "ahd"
MT evaluation scorer ended on 2013 Apr 11 at 19:56:14
```

Other Experiments Results:

For reference, here are the results of other experiments that I did along with their stats:

```
## Baseline ##
MT evaluation scorer began on 2013 Apr 7 at 22:00:24
command line:  /Users/aiman/tools/mosesdecoder/scripts/generic/mteval-v11b.pl -s data/Test/Test_
  Evaluation of Arabic-to-English translation using:
    src set "mt05_arabic_evlset_v0" (4 docs, 48 segs)
    ref set "mt05_arabic_evlset_v0-ref" (4 refs)
    tst set "mt05_arabic_evlset_v0" (1 systems)

NIST score = 6.3133  BLEU score = 0.3124 for system "ahd"

# ------------------------------------------------------------------------

Individual N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
```

```
 NIST:  5.0049   1.0157   0.2046   0.0570   0.0311   0.0108   0.0058   0.0022   0.0000  "ahd"

 BLEU:  0.7170   0.3981   0.2510   0.1582   0.0933   0.0519   0.0323   0.0169   0.0052  "ahd"

# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  5.0049   6.0206   6.2252   6.2822   6.3133   6.3241   6.3299   6.3321   6.3321  "ahd"

 BLEU:  0.6866   0.5116   0.3977   0.3124   0.2432   0.1866   0.1444   0.1098   0.0779  "ahd"
MT evaluation scorer ended on 2013 Apr 7 at 22:00:25

## Attempt 1: SVM Harmonizer Trained with 7000 Sentences ##
    - Yields dataset size: 15402 entry
    - 1972 of which belongs to class 1
    - 7000 sentences from Train_data.ar/en  (not cleaned)
    - Trained using SVM
    - Harmonized Corpus containd  2599 collapsed token
    - Harmonized Corpus size 35644 (not cleaned)
    - Test data yielded only 3 collapsed tokens

        MT evaluation scorer began on 2013 Apr 4 at 01:10:05
        command line:  /Users/aiman/tools/mosesdecoder/scripts/generic/mteval-v11b.pl -s data/Te
          Evaluation of Arabic-to-English translation using:
            src set "mt05_arabic_evlset_v0" (4 docs, 48 segs)
            ref set "mt05_arabic_evlset_v0-ref" (4 refs)
            tst set "mt05_arabic_evlset_v0" (1 systems)

        NIST score = 6.4259  BLEU score = 0.3153 for system "ahd"

        # ------------------------------------------------------------------------

        Individual N-gram scoring
                1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
                ------   ------   ------   ------   ------   ------   ------   ------   ------
         NIST:  5.0323   1.0981   0.2164   0.0523   0.0267   0.0118   0.0058   0.0046   0.0025

         BLEU:  0.7269   0.4117   0.2571   0.1556   0.0938   0.0548   0.0315   0.0190   0.0115

        # ------------------------------------------------------------------------
        Cumulative N-gram scoring
                1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
                ------   ------   ------   ------   ------   ------   ------   ------   ------
         NIST:  5.0323   6.1304   6.3468   6.3992   6.4259   6.4377   6.4435   6.4482   6.4507

         BLEU:  0.6929   0.5214   0.4054   0.3153   0.2450   0.1894   0.1456   0.1122   0.0867
        MT evaluation scorer ended on 2013 Apr 4 at 01:10:06


## Attempt 2: SVM Harmonizer Trained with 34000 Sentences without Lemmas ##

    - Yields dataset size: 11407 entry
    - 1662 of which belongs to class 1
    - 34000 sentences from Train_data.clean.ar/en
    - Trained using SVM
    - Harmonized Corpus containd 4003 collapsed token
    - Test data yielded only 7 collapsed tokens

        MT evaluation scorer began on 2013 Apr 7 at 20:45:56
        command line:  /Users/aiman/tools/mosesdecoder/scripts/generic/mteval-v11b.pl -s data/Te
          Evaluation of Arabic-to-English translation using:
            src set "mt05_arabic_evlset_v0" (4 docs, 48 segs)
            ref set "mt05_arabic_evlset_v0-ref" (4 refs)
            tst set "mt05_arabic_evlset_v0" (1 systems)

        NIST score = 6.3497  BLEU score = 0.3246 for system "ahd"

        # ------------------------------------------------------------------------

        Individual N-gram scoring
                1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
                ------   ------   ------   ------   ------   ------   ------   ------   ------
         NIST:  5.0066   1.0479   0.2163   0.0537   0.0253   0.0106   0.0066   0.0031   0.0015

         BLEU:  0.7353   0.4146   0.2733   0.1754   0.1085   0.0647   0.0402   0.0226   0.0129

        # ------------------------------------------------------------------------
        Cumulative N-gram scoring
                1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
                ------   ------   ------   ------   ------   ------   ------   ------   ------
         NIST:  5.0066   6.0545   6.2708   6.3245   6.3497   6.3603   6.3669   6.3700   6.3716

         BLEU:  0.6863   0.5154   0.4077   0.3246   0.2571   0.2020   0.1588   0.1234   0.0953
        MT evaluation scorer ended on 2013 Apr 7 at 20:45:57


## Attempt 2: SVM Harmonizer Trained with Lemmas ##
    - Lemmas are used to train the harmonizer as well
    - Yields dataset size: 11407 entry
    - 1662 of which belongs to class 1
    - 34000 sentences from Train_data.clean.ar/en
    - Trained using SVM
    - Harmonized Corpus containd 987347 non-collapsed token and 716463 collapsed token
    - Test data yielded only 7 collapsed tokens

        MT evaluation scorer began on 2013 Apr 8 at 11:41:03
        command line:  /Users/aiman/tools/mosesdecoder/scripts/generic/mteval-v11b.pl -s data/Te
          Evaluation of Arabic-to-English translation using:
            src set "mt05_arabic_evlset_v0" (4 docs, 48 segs)
            ref set "mt05_arabic_evlset_v0-ref" (4 refs)
            tst set "mt05_arabic_evlset_v0" (1 systems)

        NIST score = 6.5082  BLEU score = 0.3100 for system "ahd"

        # ------------------------------------------------------------------------

        Individual N-gram scoring
                1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
                ------   ------   ------   ------   ------   ------   ------   ------   ------
         NIST:  5.2053   1.0090   0.2110   0.0581   0.0247   0.0068   0.0027   0.0013   0.0000

         BLEU:  0.7427   0.4022   0.2418   0.1512   0.0870   0.0454   0.0237   0.0109   0.0042

        # ------------------------------------------------------------------------
        Cumulative N-gram scoring
                1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
                ------   ------   ------   ------   ------   ------   ------   ------   ------
         NIST:  5.2053   6.2144   6.4254   6.4835   6.5082   6.5150   6.5176   6.5189   6.5189

         BLEU:  0.7124   0.5242   0.3994   0.3100   0.2385   0.1796   0.1337   0.0973   0.0682
        MT evaluation scorer ended on 2013 Apr 8 at 11:41:04
```