Approaches to Machine Translation: Rule-based, Statistical and Hybrid

Language Modeling Toolkits

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Using the SRILM Toolkit
(http://www.speech.sri.com/projects/srilm/)

- **Make Counts**: Make RAW Counts from text file

```
ngram-count -text train.txt.tok.low -order 3 \
-write1 train.lm_counts.1\
-write2 train.lm_counts.2\
-write3 train.lm_counts.3
```

```
] and he       217
unto them ,    178
the son of     172
```

Using the SRILM Toolkit

- No Count-of-Counts functionality… but easy to get

```
LC_ALL=C;

cat train.lm_counts.3 \
  | awk '{print $NF}' \
  | sort -n \
  | uniq -c \
  | awk '{print $2" "$1}'
```

```
1 28054
2 6048
3 1876
4 856
```

Using the SRILM Toolkit

- **However we can obtain the GT discount factors:**

```
ngram-count -text train.tok.low\
-order 3 \
-gt1 train.gt1 -gt2 train.gt2 \
-gt3 train.gt3
```

```
mincount 1
maxcount 7
discount 1 0.548733
discount 2 0.553368
```

Using the SRILM Toolkit

- **And the Kneser-Ney:**

```
ngram-count -text train.tok.low\
-order 3 \
-kn1 train.kn1 -kn2 train.kn2 \
-kn3 train.kn3
```

```
mincount 2
discount1  0.710394
discount2  1.332489
discount3+ 1.737531
```

Using the SRILM Toolkit

- **Building the LM:**

  – **With Good-Turing:**

```
ngram-count –unk -text train.tok.low -order 3 -lm train.gt.lm
```

  – **With Witten-Bell:**

```
ngram-count –unk -text train.tok.low -order 3 -lm train.wb.lm\
        -wbdiscount
```

  – **With Unmodified Kneser-Ney:**

```
ngram-count –unk -text train.tok.low -order 3 -lm train.ukn.lm\
        -ukndiscount
```

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Using the SRILM Toolkit

- **Building the LM:**
  - **With Modified Kneser-Ney:**

```
ngram-count –unk -text train.tok.low -order 3 -lm train.kn.lm\
       –kndiscount
```

Using the SRILM Toolkit

- **Computing the perplexity:**

  - `ngram -unk –lm train.gt.lm -ppl test.tok.low`

```
file test.tok.low: 658 sentences, 19632 words, 0 OOVs
0 zeroprobs, logprob= -32504.4 ppl= 39.9935
ppl1= 45.2566
```

Using the IRSTLM Toolkit

(http://sourceforge.net/projects/irstlm/

TUTORIAL: http://www.mt-archive.info/MTMarathon-2008-Bertoldi-ppt.pdf)

- **Make Counts**: **Make RAW Counts from text file**

```
ngt -i=train.txt.tok.low \
–n=3 -gooout=y -o=train.lm_counts.3
```

```
 ] and he        217
unto them ,      178
the son of       172
```

Using the IRSTLM Toolkit

- **Count-of-Counts functionality for n1, n2, n3, n4 and n>5**

```
ngt -i=train.txt.tok.low -n=3 \
 -ikn=CC.dat
```

```
CC.dat:
```

```
level: 1  n1: 985 n2: 515 n3: 309 n4: 187 unover3: 1546
level: 2  n1: 11049 n2: 3587 n3: 1400 n4: 774 unover3: 1546
level: 3  n1: 28054 n2: 6048 n3: 1876 n4: 856 unover3: 1546
```

Using the IRSTLM Toolkit

- **Old style for n>5**

```
LC_ALL=C;

cat train.lm_counts.3 \
  | awk '{print $NF}' \
  | sort -n \
  | uniq -c \
  | awk '{print $2" "$1}’
```

```
1 28054
2 6048
3 1876
4 856
```

Using the IRSTLM Toolkit

- **(No method about obtaining WB, GT and KN statistics)**

```
But you know how to compute them from CC,
don't you?
```

Using the IRSTLM Toolkit

- **Building the Sub LM from Counts (PERL NEEDED):**

(generate 3 files: train.XX.[1-3]gr.gz)

- **With Good-Turing:**

```
build-sublm.pl --good-turing --size 3 \
                  --ngrams train.lm_counts.3 -sublm train.gt
```

- **With Witten-Bell:**

```
build-sublm.pl --witten-bell --size 3 \
                  --ngrams train.lm_counts.3 -sublm train.wb
```

- **With Unmodified Kneser-Ney:**

```
build-sublm.pl --knesser-ney --size 3 \
                  --ngrams train.lm_counts.3 -sublm train.ukn
```

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Using the IRSTLM Toolkit

- **Building the Sub LM (PERL NEEDED):**

  – **With Modified Kneser-Ney:**

```
build-sublm.pl --improved-knesser-ney --size 3 \
                --ngrams train.lm_counts.3 -sublm train.kn
```

- **Merge Sub LM to final iArpa LM format:**

```
merge-sublm.pl --size 3 --sublm train.XX --lm train.XX.lm
```

Where XX is one of the previously generated (gt, wb, ukn or kn)

Using the IRSTLM Toolkit

- **Computing the perplexity:**

  -
    ```
    bin/compile-lm train.XX.lm.gz --eval=test.tok.low
    ```

```
%% Nw=12290 PP=116.32 PPwp=34.58 Nbo=7135 Noov=269
OOV=2.19%
```