Karlsruhe Institute of Technology
Elizabeth Salesky
Institute for Anthropomatics and Robotics, Interactive Systems Labs
elizabeth.salesky@kit.edu

MT Praktikum - Segmentation

10. Juli 2018

Environment Setup

First we need to have access to our cluster enviroment.
If you can use the available machines in the pool room, please log in with username: smt[30-45] (any number between 30 to 45, for example smt35), password=123456.

If you use your own laptop, you can directly connect to the cluster using ssh, using this command:

```
ssh smt[30-45]@i13hpc1.ira.uka.de
```

Next, please log into i13hpc28 or i13hpc29, using the following commands:

```
ssh i13hpc1 (this is our login server; if you use your laptop, you are already here)
ssh i13hpc28 or ssh i13hpc29
```

From there, go to your working directory:

```
cd /project/smtstud/ss18/systems/{username}/
```

Now enter the pre-installed virtual environment using these 2 commands:

```
bash
. /project/smtstud/ss18/commands/setup.sh
```

If you see the (praktikum) at the beginning of your terminal line, the setup was successful.

This directory contains today's data and scripts. Copy this directory into your working directory:

```
/project/smtstud/ss18/data/seg
```

NMT: Vocabulary

We are going to compare segmentation schemes for different languages. You have `train`, `dev`, and `test` data for English (en), German (de), and Turkish (tr). Log into rg3hpc1, and then to i13hpc28 or i13hpc29. Once you are in either i13hpc28 or i13hpc29, move into your working directory.

First, we will look at vocabulary sizes and the out-of-vocabulary problem.

1. For each language, calculate the vocabulary size of the training data (number of unique words in `train.x`).
   - English:
   - German:
   - Turkish:
2. For each language, what is the average sentence length for the training data?
   - English:
   - German:
   - Turkish:
3. How many words occur only 1× in `train.en`? Only 5×?

   How many words occur only 1× in `train.de`? Only 5×?

   How many words occur only 1× in `train.tr`? Only 5×?
4. A common scheme for word-based NMT is to take the 50k most common words as the vocabulary. This can lead to out of vocabulary words in dev and test. If you take the top 50k words from `train` as your vocabulary, how many out of vocabulary words are in `dev`?
   - English:
   - German:
   - Turkish:

   And in `test`?
   - English:
   - German:
   - Turkish:

Byte-Pair Encoding (BPE)

To get around the out-of-vocabulary problem, many people use BPE. Now, you will use scripts from `subword-nmt` to train BPE. BPE iteratively joins together common character sequences, creating subwords. The number of operations is the number of merges it creates.

To train BPE units, use the following command:

```
./subword-nmt/learn_bpe.py -s {num_operations} < {train_file} > {codes_file}
```

To apply the BPE codes to a file, use the following command:

```
./subword-nmt/apply-bpe -c {codes_file} < {in_file} > {out_file}
```

For these exercises, codes have been created for you to save time, using 20k and 50k merge operations. They are in `codes/`.

1. The 20k codes files have been applied to `train` for you, creating files named `bped/train.x.20k`.

   What is the vocabulary size of the new BPE-d training data now?

   - English:
   - German:
   - Turkish:

2. Now, what is the average sentence length for the training data?

   - English:
   - German:
   - Turkish:

   On average, this means that there are how many subword units per word?

   - English:
   - German:
   - Turkish:

3. Why do you think the change is different for different languages?

4. Apply the codes to `dev` and `test` using the command above. Using the new training data as vocabulary, how many OOVs are in `dev` and `test`?

   - English:
   - German:
   - Turkish:

5.

6. The 50k codes files have been applied to `train` for you, creating files named `bped/train.x.50k`.

   Looking at these files, what is the vocabulary size of the new BPE-d training data now?

- English:
- German:
- Turkish:

7. Now, what is the average sentence length for the training data?

   - English:
   - German:
   - Turkish:

   On average, this means that there are how many subword units per word?

   - English:
   - German:
   - Turkish:

   How is this different than with 20,000 operations?

   (Notice, the first 20,000 operations in the 50,000 codes files are the same as the 20,000 codes files).

8. Apply the 50k codes to `dev` and `test`. Using the new training data as vocabulary, how many OOVs are in `dev` and `test`? Why do you think they're different than with 20k, if they are?

   - English:
   - German:
   - Turkish:

Language Modeling
If you have time, you can look at how segmentation affects model perplexity, using n-gram language modeling.

To train an order-n language model, use this command:

```
lmplz -o {order} < {training_data} > {lm_name.arpa}
```

To get the perplexity of a dataset using an arpa file, use this command:

```
python perp.py {arpa_file} {text_data}
```

(this may take a minute or so to execute, depending on the size of the LM and dataset)
For these exercises, LMs for **German** have been trained for you to save time (in `/project/smtstud/ss18/data/lms/`), but you're encouraged to look at the other two languages, or try different order LMs, if you have time. You do NOT need to copy these LMs to your working directory; you can use a relative or absolute path to the arpa file, or softlink (`ln -s`) it in your working directory if that's easier for you.

1. First, look at the order-3 LM in `lms/` for the original data (i.e. not BPE-ed). For each language, use these LMs to calculate the perplexity of the `dev` data. What is the perplexity?

2. Now, use the order-3 LM in `lms/` for the 20k BPE data to calculate the perplexity of the `dev` data. What is the perplexity? Has it changed from using words?

3. Finally, use the order-3 LM in `lms/` for the 50k BPE data to calculate the perplexity of the `dev` data. What is the perplexity? Has it changed from the previous two?

4. How do you think this would change if you used a larger order (e.g. 5) LMs?