LIN570: HW4 – morphological segmentation (2/2) (100pts)

YOUR NAME (`UW NetID`)

Due date: 11pm on Oct 29, 2018 (Tuesday)

For this homework, you are going to reproduce compound splitting (Koehn and Knight, 2003). All the required files are under `~/dropbox/19-20/570/hw4/examples`.

Rubric:

2pts `hw.tar.gz` submitted, it should contain following files:

- `translation_lexicon.sh` for Q1
- `translation.lexicon.proba.head` for Q1
- `file.segmented` for Q1
- `segment_german_corpus.sh` for Q2
- `second.translation.lexicon.proba.head` for Q2
- `file.second.segmented` for Q2

2pts `readme.txt` or `readme.pdf` submitted

6pts All files and folders are present in expected locations

10pts Programs run to completion

5pts The output of programs on patas match submitted output

To display utf8 characters correctly in patas:

```
export LC_ALL=en_US.UTF-8
export LANG=en_US.UTF-8
export LANGUAGE=en_US.UTF-8
```

1. it requires HW3 Q2 (frequency based metric)
2. (50pts) guidance from a parallel corpus, which requires a translation lexicon from a parallel corpus. This can be done with a toolkit `giza++` (Och and Ney, 2003)[1]

   DE → EN

   ```
   # Sentence pair (1) source length 4 target length 3 alignment score : 0.00540133
   Wiederaufnahme der Sitzungsperiode
   NULL ({ }) resumption ({ 1 }) of ({ }) the ({ 2 }) session ({ 3 })
   ```

[1] `https://github.com/moses-smt/giza-pp`

$Wiederaufnahme_1$ der_2 $Sitzungsperiode_3$

resumption	$Wiederaufnahme_1$
the	der_2
session	$Sitzungsperiode_3$

```
# Sentence pair (575) source length 7 target length 10 alignment score : 1.44592e-30
mir scheint , daß eher das Gegenteil richtig ist .
NULL ({ 3 9 }) I ({ 1 }) find ({ 2 }) the ({ 4 }) opposite ({ 5 6 7 })
the ({ }) case ({ 8 }) . ({ 10 })
```

mir_1 $scheint$ $,_3$ $daß_4$ $eher_5$ das_6 $Gegenteil_7$ $richtig_8$ ist_9 $._{10}$

NULL	$,_3$ ... ist_9
I	mir_1
find	*scheint*
...	
opposite	$eher_5$ das_6 $Gegenteil_7$
...	
.	$._{10}$

EN → DE

```
# Sentence pair (1) source length 3 target length 4 alignment score : 0.000276873
resumption of the session
NULL ({ }) Wiederaufnahme ({ 1 }) der ({ 2 }) Sitzungsperiode ({ 3 4 })
```

$resumption_1$ of_2 the_3 $session_4$

Wiederaufnahme	$resumption_1$
der	of_2
Sitzungsperiode	the_3 $session_4$

```
# Sentence pair (575) source length 10 target length 7 alignment score : 2.72174e-14
I find the opposite the case .
NULL ({ }) mir ({ 1 }) scheint ({ 2 }) , ({ }) daß ({ }) eher ({ }) das ({ 3 })
Gegenteil ({ 4 5 }) richtig ({ 6 }) ist ({ }) . ({ 7 })
```

I_1 $find_2$ the_3 $opposite_4$ the_5 $case_6$ $._7$

NULL	
mir	I_1
...	
Gegenteil	$opposite_4$ the_5
...	

(a) (30pts) implement a script to make the translation lexicon (de → en) from the GIZA++ result and to calculate their probabilities (`translation.lexicon.proba`) with a threshold 0.01

```
768 Dienstleistungsrichtlinie       Services Directive
225 Dienstleistungsrichtlinie       services directive
 34 Dienstleistungsrichtlinie       directive ... services
  6 Dienstleistungsrichtlinie       services
  6 Dienstleistungsrichtlinie       Directive ... Services
  5 Dienstleistungsrichtlinie       Services ... Directive
  5 Dienstleistungsrichtlinie       service directive
  4 Dienstleistungsrichtlinie       Directive ... services
  3 Dienstleistungsrichtlinie       Service Directive
  3 Dienstleistungsrichtlinie       directive
```

`translation.lexicon.proba` exmaple:

```
Dienstleistungsrichtlinie       Services Directive (0.725212465); services
directive (0.212464589); directive ... services (0,03210576)
```

where the translation probability of *services* is 0.005665722 (<0.01).

- `zcat en-de.A3.final.gz | ./translation_lexicon.sh > translation.lexicon.proba`
- `sort translation.lexicon.proba | head > translation.lexicon.proba.head`

(b) (20pts) implements a script to find a best segmentation using a translation lexicon table. (`file.segmented`)

- for each German word, we consider all splitting options
- for each splitting option, we check if it has translations on the English side
- we allow each English word to be considered only once: if it is taken as evidence for correspondence to the first part of the compound, it is excluded as evidence for the other parts.
- if multiple options match the English, we select the one(s) with the most splits and use word frequencies

3. (25pts) second translation table

while works well for *Aktionsplan* and *Freitag* using solution #1, fails for *Grundrechte* ('basic right').

- *Grundrechte* ('basic rights'): *Grund* + *Rechte*
- *Grund* is translated into (N. 'reason' or 'foundation') → we are looking for (ADJ. 'basic' or 'fundamental')

Solution 2 (alignment with split compounds)

Die Charta der ***Grundrechte*** *der Europäischen Union*

↓

Die Charta der ***Grund rechte*** *der Europäischen Union*

⇕

The Charter of Fundamental Rights of the European Union

(a) (10pts) convert `europarl-v7.de-en.lowered.de` into `europarl-v7.de-en.lowered-segmented.de`:
`cat europarl-v7.de-en.lowered.de | ./segment_german_corpus.sh > europarl-v7.de-en.lowered-s`

(b) (15pts) run a word align and repeat Q1 to produce `second.translation.lexicon.proba.head` and (`file.second.segmented`)

```
~/mosesdecoder/scripts/training/clean-corpus-n.perl \
    europarl-v7.de-en.lowered-segmented de en \
    europarl-v7.de-en.clean 1 80
```

```
~/mosesdecoder/scripts/training/train-model.perl -root-dir train  \
-corpus europarl-v7.de-en.clean -f de -e en \
-alignment  grow-diag-final-and -reordering msd-bidirectional-fe \
-lm 0:3:$PWD/news-commentary-v8.fr-en.blm.en:8 \
-external-bin-dir ~/mosesdecoder/tools \
--first-step 1 --last-step 4
```

(c) NOTE:

- you don't need to compile `moses`
- you need scripts in `~/mosesdecoder/scripts`
- you also need GIZA++'s binaries in `~/mosesdecoder/tools`: GIZA++ MKCLS PLAIN2SNT.OUT SNT2COOC.OUT SNT2PLAIN.OUT
- a dummy lm file is provided

```
(1) prepare corpus
(2) run GIZA
(3) align words
(4) learn lexical translation
(5) extract phrases
(6) score phrases
(7) learn reordering model
(8) learn generation model
(9) create decoder config file
```

```
corpus:
total 702032
-rw-rw-r-- 1 jungyeul jungyeul 346182970 Oct  9 13:16 de-en-int-train.snt
-rw-rw-r-- 1 jungyeul jungyeul   9099967 Oct  9 13:15 de.vcb
-rw-rw-r-- 1 jungyeul jungyeul   6792426 Oct  9 13:03 de.vcb.classes
-rw-rw-r-- 1 jungyeul jungyeul   5671637 Oct  9 13:03 de.vcb.classes.cats
-rw-rw-r-- 1 jungyeul jungyeul 346182970 Oct  9 13:18 en-de-int-train.snt
-rw-rw-r-- 1 jungyeul jungyeul   2201432 Oct  9 13:15 en.vcb
-rw-rw-r-- 1 jungyeul jungyeul   1523858 Oct  9 13:15 en.vcb.classes
-rw-rw-r-- 1 jungyeul jungyeul   1176249 Oct  9 13:15 en.vcb.classes.cats

giza.de-en:
total 1281864
-rw-rw-r-- 1 jungyeul jungyeul 358877908 Oct 10 00:06 de-en.A3.final.gz
```

```
-rw-rw-r-- 1 jungyeul jungyeul 953723845 Oct  9 13:29 de-en.cooc
-rw-rw-r-- 1 jungyeul jungyeul      1835 Oct  9 13:30 de-en.gizacfg

giza.en-de:
total 1843768
-rw-rw-r-- 1 jungyeul jungyeul 359279141 Oct 10 10:00 en-de.A3.final.gz
-rw-rw-r-- 1 jungyeul jungyeul 951358474 Oct 10 00:17 en-de.cooc
-rw-rw-r-- 1 jungyeul jungyeul      1835 Oct 10 00:18 en-de.gizacfg
```

References

Koehn, P. and Knight, K. (2003). Empirical Methods for Compound Splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pages 187–194.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.