Exercise 9

Evgeniia Tokarchuk 383433
Petre Petrov 383349
Aman Gokrani 383477

June 2018

1

First of all, obtain the most frequent translation word by word (assume here direct alignment) using python.

```
at-voon {'ok-voon': 3}
bichat {'ororok': 2}
dat {'sprok': 5}
. {'.': 11, 'zanzanok': 1}
at-drubel {'ok-drubel': 1}
pippat {'anok': 2, 'drok': 1}
rrat {'plok': 2}
totat {'erok': 1, 'wiwok': 2}
arrat {'izok': 2, 'crrrok': 1}
vat {'hihok': 2, 'izok': 1}
hilat {'ghirok': 1, 'clok': 1}
krat {'anok': 1, 'izok': 1}
sat {'brok': 1}
lat {'jok': 1, 'brok': 1}
jjat {'farok': 2}
quat {'izok': 2, 'jok': 1}
cat {'stok': 2}
wat {'lalok': 6}
eneat {'enemok': 1}
iat {'lalok': 1}
nnat {'nok': 3, 'mok': 1, 'rarok': 1}
oloat {'kantok': 1}
at-yurp {'ok-yurp': 1}
gat {'nok': 1, 'mok': 1}
mat {'yorok': 1, 'hihok': 1}
bat {'ghirok': 1}
zanzanat {'yorok': 1}
forat {'nok': 1}
```

from this obtain initial vocabulary, assuming that words, which has only one translation and this translation was done more than once, it will be translation of this word. Also from statistics we can do some more conclusions: nnat will be translated to nok (because of highest rate of translations) and at-drubel → ok-drubel, at-yurp → ok-yurp by similarity.

Table 1: Initial vocabulary

Arcturian	Centauri
at-voon	ok-voon
bichat	ororok
dat	sprok
rrat	plok
jjat	farok
cat	stok
wat	lalok
nnat	nok
at-yurp	ok-yurp
at-drubei	ok-drubei

Now we can go step by step through sentences and find intersection among possible translations.

1. Sentence 2:
pippat → anok (only one word not in the initial vocabulary, so translation obtained directly)
2. Sentence 3:
otat → erok, izok, hihok,ghirok
arrat → erok, izok, hihok,ghirok
vat → erok, izok, hihok,ghirok
hilat → erok, izok, hihok,ghirok

3. Sentence 4:
krat → drok, brok, jok
sat → drok, brok, jok
lat → deok, brok, jok

4. Sentence 5:
totat → wiwok or izok
quat → wiwok or izok

5. Sentence 6:
krat → izok or jok: from sentence 4 and 6 krat = jok
quat → izok, jok: according to previous obtained translation quat=izok

6.Return back to sentence 4 because of new vocabulary
sat – ¿ drok, brok

lat → drok, brok

7. Sentence 7
vat → izok, enemok: from sentence 2 and current sentece 7 vat=izok
eneat → izok, enemok: eneat=enemok (because of vat translation)

8. Update sentence 3 according new translations:
totat → erok, hihok, ghirok
arrat → erok, hihok, ghirok
hilat → erok, hihok, ghirok

9. Sentence 8:
iat → lalok, brok
lat → lalok, brok: from sentence 4 and current lat=brok ==¿ iat=lalok

10. Sentence 9:
totat → wiwok, kantok: from sentence 5 and current totat=wiwiok
oloat → wiwok,kantok from previous ==¿ oloat=kantok

11. Sentence 10:
gat → mok, yorok, ghirok, clok
mat → mok, yorok, ghirok, clok
bat → mok, yorok, ghirok, clok
hilat → mok, yorok, ghirok, clok: from sentence 3 by intersection hilat=ghirok

12. Update sentence 3 with according new vocab:
totat → erok, hihok
arrat → erok, hihok

13. Sentence 11:
arrat → crrrok, hihok, yorok, zanzanok: from sentence 3 arrat=hihok ==¿ totat=erok
mat → crrrok, yorok, zanzanok (sent 10 - yorok)
zanzanat → crrrok, yorok, zanzanok

14. Sentence 12:
forat → rarok, mok
gat → rarok, mok: From sentence 10 gat=mok ==¿ forat=rarok

15. Update sentence 10:
bat = clock

16. Last translation
zanzanat → crrrok, zanzanok (by posotion assign zananat=zanzanok)

From this produce final vocabulary:

Table 2: Final vocabulary

Arcturian	Centauri
at-voon	ok-voon
bichat	ororok
dat	sprok
rrat	plok
jjat	farok
cat	stok
wat	lalok
nnat	nok
at-yurp	ok-yurp
at-drubel	ok-drubel
pippat	anok
krat	jok
quat	izok
lat	brok
iat	lalok
vat	izok
eneat	enemok
totat	wiwok, erok
oloat	kantok
hilat	ghirok
arrat	hihok
mat	yorok
gat	mok
forat	rarok
bat	clock
zanzanat	zanzamok

Translation according the vocabulary:

1. direct: lalok brok anok enemok ghirok kantok ok-yurp
   There is no bigrams (enemok, ghirok) and (ghirok, kantok) =¿ change the order of enemok and ghirok
   Result: lalok brok anok ghirok enemok kantok ok-yurp

2. direct: wiwok/erok nok rarok hihok yorok clock
   From bigrams result: wiwok rarok nok hihok yorok clock

3. direct: lalok sprok izok stok ___ ok-drubel
   The missing word obtained from bigrams: vok
   Result: lalok sprok izok stok vok ok-drubel

2

The error rates implemented in error_rates.py python file.

With punctuation (average error rate):
WER: 0.4778153295397505
PER: 0.4497648514271587

Without punctuation (average error rate):
WER: 0.5142105431874499
PRE: 0.4487520419983171

3

(a)

IBM Model 2 addresses the issue of alignment with an explicit model for alignment based on the positions of the input and output words. The translation of a foreign input word in position i to an English word in position j is modeled by an alignment probability distribution

(b)

(c)

Add fertility model. Fertility of input words is modeled directly with a probability distribution $p(\phi|f)$.
For each foreign word f, this probability distribution indicates how many $\phi = 0, 1, 2, ...$ output words it usually translates to.

(d)

In IBM Model 4, each word is dependent on the previously aligned word and on the word classes of the surrounding words. That means that some words trigger reordering and creates a condition for how the reordering should be made

4

The implementation is done in ibm1.py Top 30 english words with most probable translations (with vocabulary size 10000 woeds):

the	die 0.16	UNK 0.15	der 0.13
,	, 0.23	UNK 0.21	die 0.07
	. 0.27	UNK 0.16	die 0.09
of	UNK 0.22	der 0.14	, 0.09
to	, 0.18	zu 0.17	UNK 0.13
and	und 0.49	UNK 0.13	, 0.07
in	in 0.299	UNK 0.1405	, 0.07
a	UNK 0.16	eine 0.15	ein 0.12
that	dass 0.28	, 0.22	die 0.09
is	ist 0.39	. 0.09	, 0.08
@-@	UNK 0.29	@-@ 0.10	, 0.09
's	UNK 0.24999	der 0.15	die 0.09
for	für 0.39	UNK 0.13	die 0.08
"	" 0.72	UNK 0.08	, 0.04
-	- 0.75	UNK 0.05	die 0.02
it	es 0.28	, 0.13	sie 0.11
as	wie 0.20	als 0.19	UNK 0.12
be	werden 0.18	sein 0.17	, 0.10
on	auf 0.25	UNK 0.15	, 0.09
are	sind 0.36	, 0.09	UNK 0.098
with	mit 0.48	UNK 0.11	, 0.06
but	aber 0.39	doch 0.19	allerdings 0.06
by	durch 0.20	UNK 0.15	von:0.10
not	nicht 0.63	sondern 0.05	, 0.04
has	hat:0.43	UNK 0.07	. 0.07
have	haben 0.37	UNK 0.08	, 0.08
this	dies 0.16	diese 0.13	dieser 0.11
will	wird 0.41	werden 0.19	, 0.06
from	von 0.20	aus 0.20	UNK 0.13
its	seine 0.18	seiner 0.09	ihre 0.08