Lehrstuhl für Informatik 6
Rheinisch–Westfälische Technische Hochschule Aachen
Prof. Dr.–Ing. H. Ney

June 16, 2015
A. Guta, M. Nuhn

8. Exercise Sheet **Statistical Methods in Natural Language Processing**

The solutions to the problems may be submitted until **Tuesday, June 23, 2015** before the exercise lesson. Upload a digital version to the L2P and bring your written/printed solution to the exercise lesson. Condition for obtaining the **Leistungsnachweis** (*Schein*) "Statistical Methods in Natural Language Processing" is the successful solution of 50% of the problems and the presentation of the solution of at least two problems in the exercise lessons.
The solutions to the problems can be submitted in groups of up to **two** students.
For programming exercises:

- The implementation has to be done in C++, but you can use additional standard Unix tools (tr, sed, awk...) for the preprocessing.
- Implementations have to be uploaded to L2P as well as printed out and submitted before the exercise lesson.
- Upload the programs as a `.tgz` or a `.zip` compressed directory to L2P. It must include a `Makefile` and must compile on Linux using simply the `make` command.
- Include a *short* description of the main data structures and algorithms used in your solution sheet.

1. [5 Points] Improve the spelling correction algorithm from the previous exercise.

   a) [2 Points] Use a higher order language model to do the spelling correction. How do the dynamic programming equations change? What error rate reduction do you obtain?

   b) [3 Points] Implement the minimum symbol error rate decision rule

   $$\hat{c}_n = \arg\max_c \{ \sum_{c_1^N : c_n = c} \Pr(c_1^N | x_1^N) \}$$

2. [1 Point] Select five sentences and submit them to an online translation service. Translate them from English to another language and back to English. Rate the resulting sentences for grammaticality and preservation of meaning. Repeat the process; does the second round of iteration give worse results or the same results? Does the choice of intermediate language make a difference to the quality of the results?

3. [1 Point] The IBM Model 3 machine translation model assumes that, after the word choice model proposes a list of words and the offset proposes possible permutations of the words, the language model can choose the best permutation. This exercise investigates how sensible that assumption is. Try to unscramble these proposed sentences into the correct order:

   - have programming a seen never I language better
   - loves john mary
   - is the communication exchange of intentional information brought by about the production perception of and signs from drawn a of system signs conventional shared

   Which ones could you do? What type of knowledge did you draw upon? Additionally you can train an n-gram model from a training corpus (you can use a publicly available toolkit, e.g. the SRI toolkit), and use it to find the highest-probability permutation of some sentences from a test corpus. Report on the accuracy of this model.

4. [3 Points] Suppose you want to translate following sentences from Arcturan to Centauri:

   a) iat lat pippat eneat hilat oloat at-yurp .

   b) totat nnat forat arrat mat bat .

   c) wat dat quat cat uskrat at-drubel .

   In order to accomplish this, you are given following bilingual text:

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

   You are also given this additional monolingual Centauri text:

   ok-drubel anok ghirok farok . wiwok rarok nok zerok ghirok enemok . ok-drubel ziplok stok vok erok enemok kantok ok-yurp zinok jok yorok clok . lalok clok izok vok ok-drubel . ok-voon ororok sprok . ok-drubel ok-voon anok plok sprok . erok sprok izok hihok ghirok . ok-voon anok drok brok jok . wiwok farok izok stok . lalok sprok izok jok stok . lalok brok anok plok nok . lalok farok ororok lalok sprok izok enemok . wiwok nok izok kantok ok-yurp . lalok mok nok yorok ghirok clok . lalok nok crrrok hihok yorok zanzanok . lalok rarok nok izok hihok mok .

   with the corresponsing bigram counts

1	. erok	1	hihok yorok	1	ok-drubel ok-voon	7	. lalok	1	izok enemok
1	ok-drubel ziplok	2	. ok-drubel	2	izok hihok	2	ok-voon anok	2	. ok-voon
1	izok jok	1	ok-voon ororok	3	. wiwok	1	izok kantok	1	ok-yurp .
1	anok drok	1	izok stok	1	ok-yurp zinok	1	anok ghirok	1	izok vok
1	ororok lalok	2	anok plok	1	jok .	1	ororok sprok	1	brok anok
1	jok stok	1	plok nok	1	brok jok	1	jok yorok	1	plok sprok
2	clok .	2	kantok ok-yurp	2	rarok nok	1	clok izok	1	lalok brok
2	sprok .	1	crrrok hihok	1	lalok clok	3	sprok izok	1	drok brok
1	lalok farok	2	stok .	2	enemok .	1	lalok mok	1	stok vok
1	enemok kantok	1	lalok nok	1	vok erok	1	erok enemok	1	lalok rarok
1	vok ok-drubel	1	erok sprok	2	lalok sprok	1	wiwok farok	1	farok .
1	mok .	1	wiwok nok	1	farok izok	1	mok nok	1	wiwok rarok
1	farok ororok	1	nok .	1	yorok clok	1	ghirok .	1	nok crrrok
1	yorok ghirok	1	ghirok clok	2	nok izok	1	yorok zanzanok	1	ghirok enemok
1	nok yorok	1	zanzanok .	1	ghirok farok	1	nok zerok	1	zerok ghirok
1	hihok ghirok	1	ok-drubel .	1	zinok jok	1	hihok mok	1	ok-drubel anok
1	ziplok stok								

   Give what you think are the most probable translations of the three sentences listed above. Explain shortly how you arrive at them.