Machine Translation
WiSe 2015/2016

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Exercise 4

Dr. Mariana Neves

December 14th, 2015

Exercises

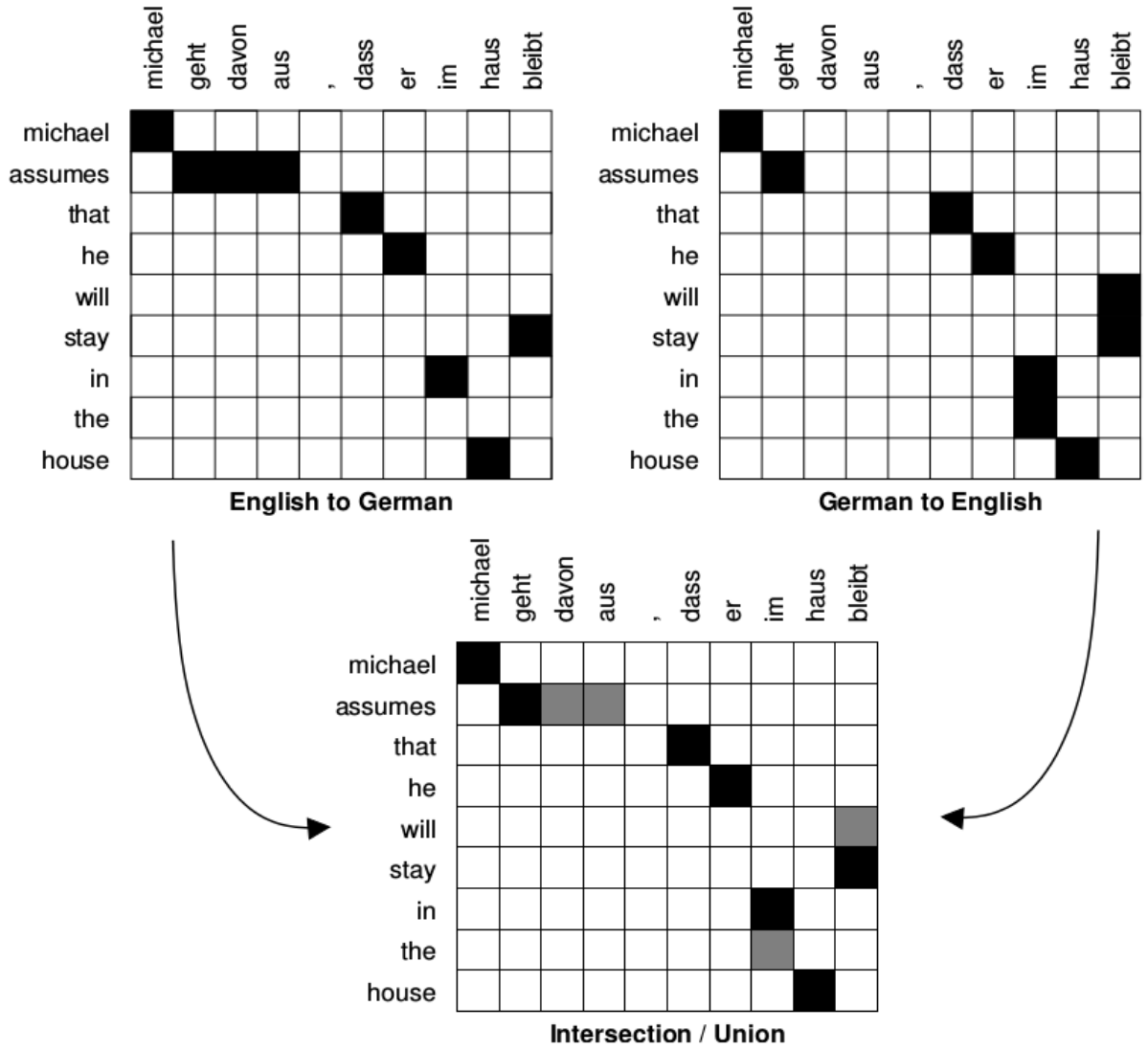
1. Word alignment
2. Language models

Optionally, also the improvement of the IBM model(s)

Submission

- Deadline
  - Sunday, Jan 31st 2016, 23:59
- Hand-in
  - SQL file with queries
  - Stored procedures in the team's schema
  - Python/Java code
- Presentation
  - Monday, Feb 1st 2016

Word alignment



grow-diag-final(e2f,f2e)

1: neighboring = {(-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1)}
2: alignment A = intersect(e2f,f2e); grow-diag(); final(e2f); final(f2e);

grow-diag()

1: **while** new points added **do**
2: **for all** English word $e \in [1...e_n]$, foreign word $f \in [1...f_n]$, $(e, f) \in A$ **do**
3: **for all** neighboring alignment points (e_{new}, f_{new}) **do**
4: **if** (e_{new} unaligned OR f_{new} unaligned) AND (e_{new}, f_{new}) $\in$ union(e2f,f2e) **then**
5: add (e_{new}, f_{new}) to A
6: **end if**
7: **end for**
8: **end for**
9: **end while**

final()

1: **for all** English word $e_{new} \in [1...e_n]$, foreign word $f_{new} \in [1...f_n]$ **do**
2: **if** (e_{new} unaligned OR f_{new} unaligned) AND (e_{new}, f_{new}) $\in$ union(e2f,f2e) **then**
3: add (e_{new}, f_{new}) to A
4: **end if**
5: **end for**

Word alignment

- Based on your implementation of the IBM model 1 or 2
- Training of the models on any possible number of sentences
- Evaluation:
  - Manual validation on a sample of 10 sentences from the corpus (not included in the training data)

Language model

- Implementation of a bigram language model
  - Including add-one smoothing
- For both languages: English and German
- Integration into the IBM model(s)

$$\begin{aligned} \operatorname{argmax}_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) &= \operatorname{argmax}_{\mathbf{e}} \frac{p(\mathbf{f}|\mathbf{e})\ p(\mathbf{e})}{p(\mathbf{f})} \\ &= \operatorname{argmax}_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})\ p(\mathbf{e}) \end{aligned}$$