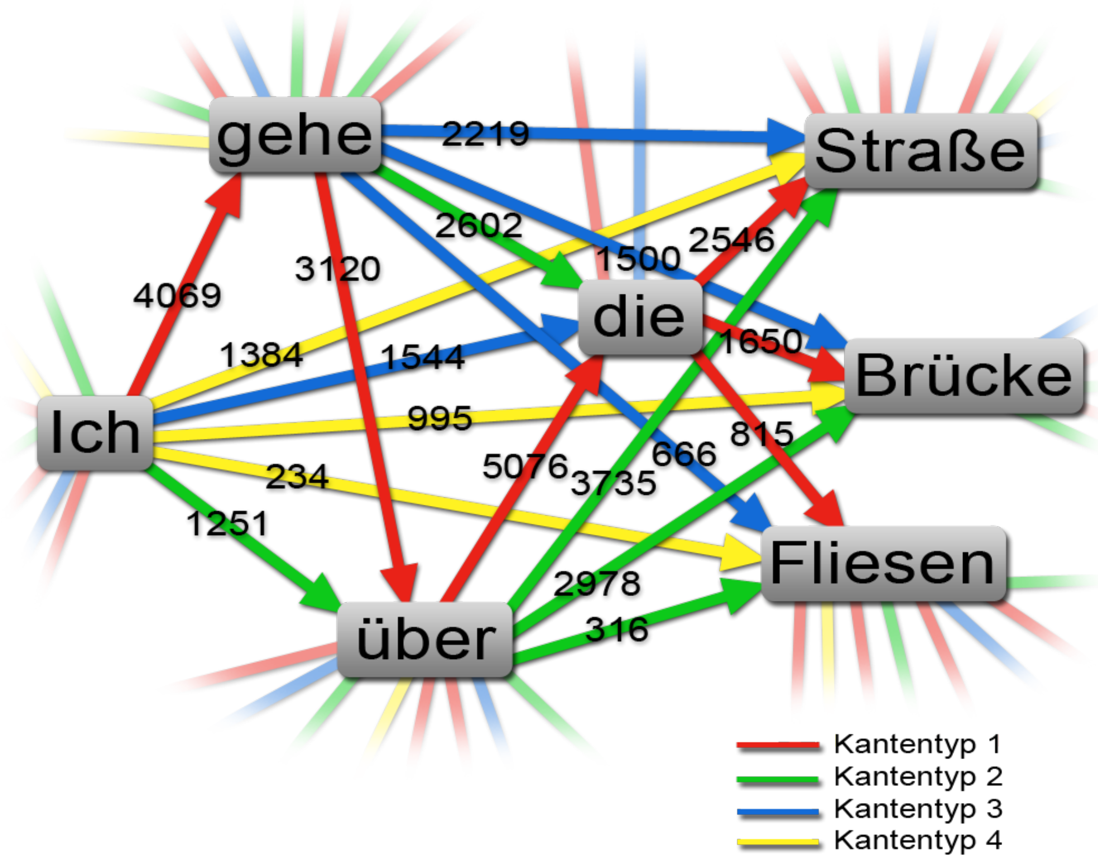
Typology: A graph based method for sentence completion



by René Pickhardt, Dr. Thomas Gottron, Paul Wagner and Till Speicher

this is work in progress

Outline of this talk

- Live Demonstration
- Discussion of the preliminary results (Motivation)
- Overview of the related work
- Introduction of our (yet informal) model
- An overview for the planned evaluation
- Outlook / Applications

What German Scentence do I want to write?

(11 Keystrokes)

- Ich
  b a e Mi d l H

(15 to 18 Keystrokes)

- Darf
  i I e Tas K anbi

Ich bin auch ein Mitglied der letzten Hälfte (37 Keystrokes)

- Typology
  - Ich b a e Mi d l H (11 Keystrokes)
- Unigrams (T9)
  - Ich bin auc ein Mitglied d le Hä (21 - 25)

Darf ich Ihnen eine Tasse Kaffe anbieten? (34 Keystrokes)

- Typology
  - Darf i I e Tas K anbi (15 - 18)
- Unigrams (T9)
  - Darf ic Ih ei Tass Kaff anbi (22)

Federal Competition Jugend Forscht:

- 4th place out of 457 projects in their field
- Special award by Gesellschaft fuer Informatik
- Invitation to International Science Fair in Washington

Google Science Fair:

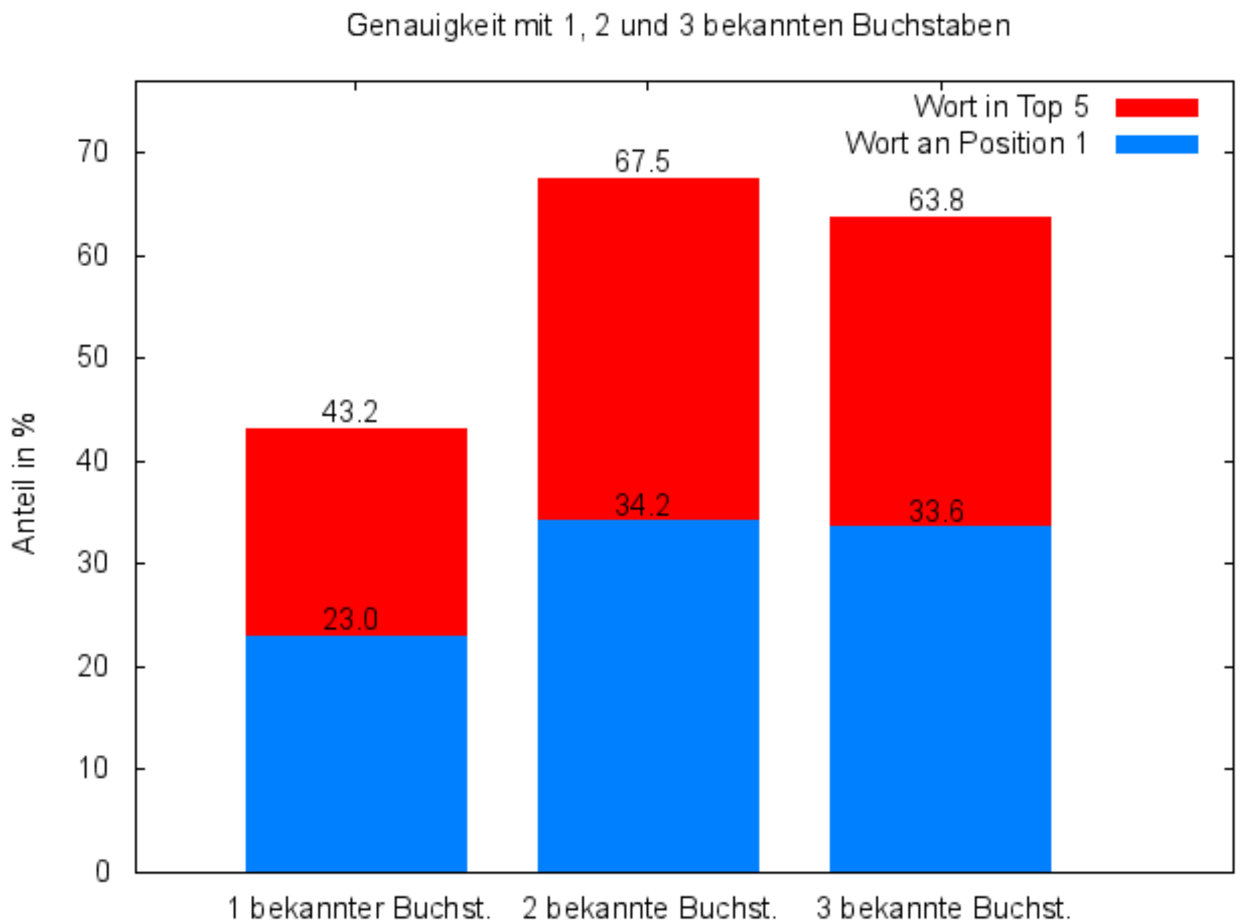
- awarded top 90 projects world wide
  - out of over 1000 submissions

Paul Wagner   Rene Pickhardt   Till Speicher

Outline of this talk

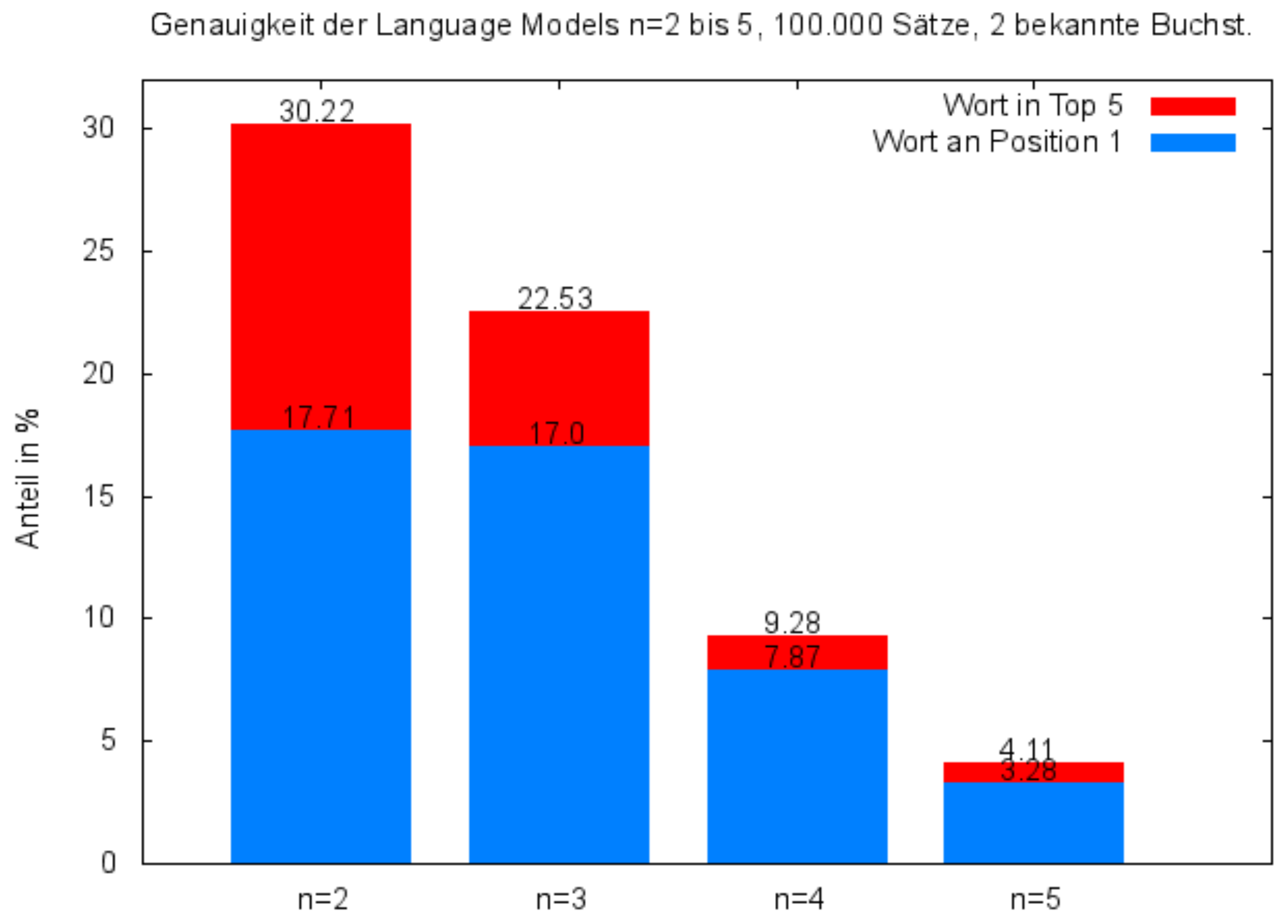
- Live Demonstration
- Discussion of the preliminary results (Motivation)
- Overview of the related work
- Introduction of our (yet informal) model
- An overview for the planned evaluation
- Outlook / Applications

We measured the Quality of the Results



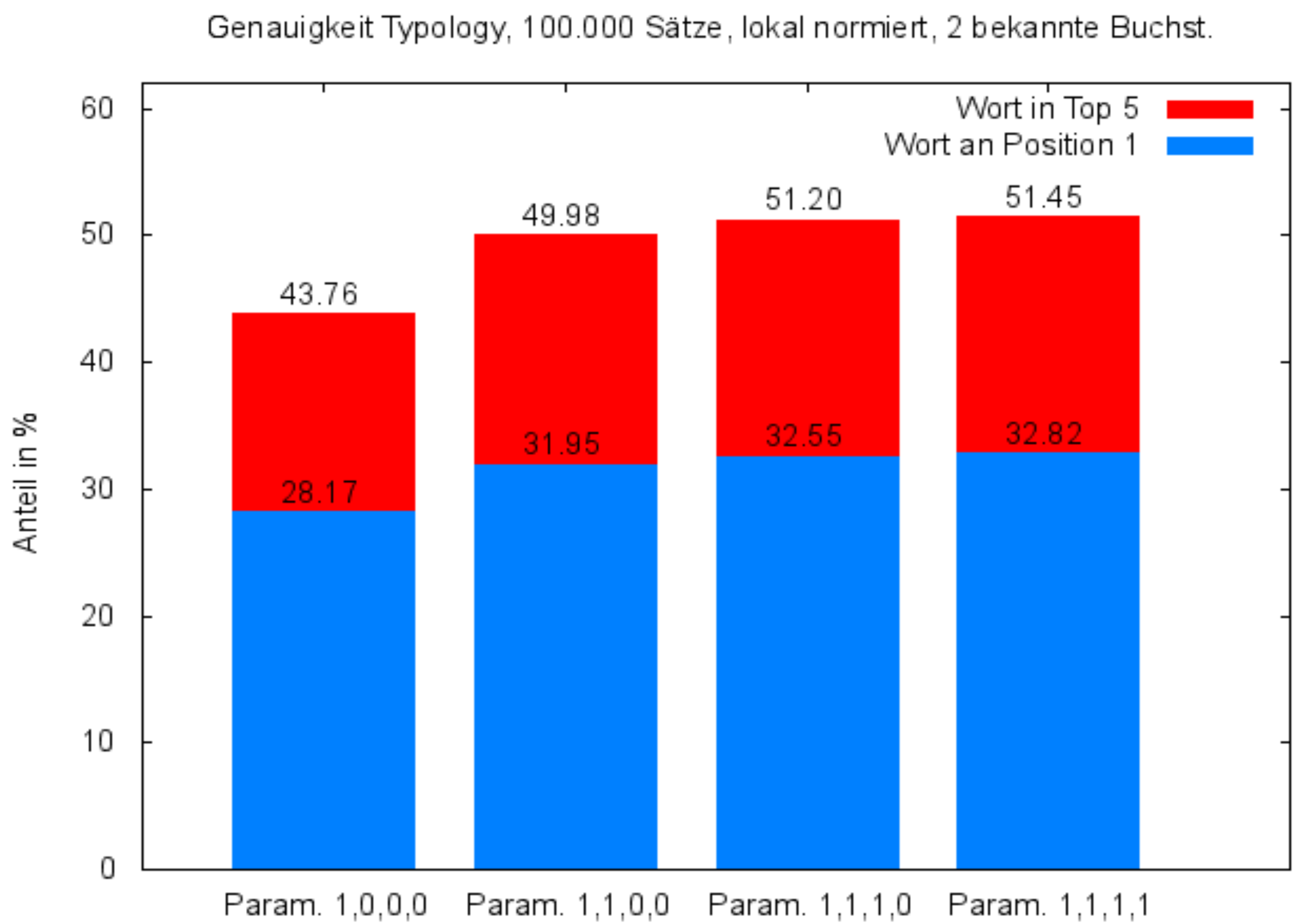
- if 2 letters of the next word are known
- in 2 out of 3 times the correct word is in the top 5 suggestions
- Is this a high Quality?

What about standard methods (Language Models)



- Language Models don't achieve results in a close range
- even worse: increasing the length of the query predictions become worse
- Data sparsity

Typology gains accuracy with longer queries



- Typology obviously has less problems with sparse data

Outline of this talk

- Live Demonstration
- Discussion of the preliminary results (Motivation)
- Overview of the related work
- Introduction of our (yet informal) model
- An overview for the planned evaluation
- Outlook / Applications

Widely studied problem

- Language Models
  - 1998 Ponte et al. (later: 2001 SIGIR, SIGMOD 2001)
- Query Prediction
  - 2006 SIGIR Holger Bast, 2011 VLDB, 2011 WWW
- Text Prediction
  - 2004 SIGIR, 2005 ECML, 2007 VLDB, 2010 ECIR
- Graph Mining
  - 2003 Schenker et al.
- Spreading activation
  - 1975 psychology
  - 2005 IEEE Boosting item keyword search with Sp. Act.

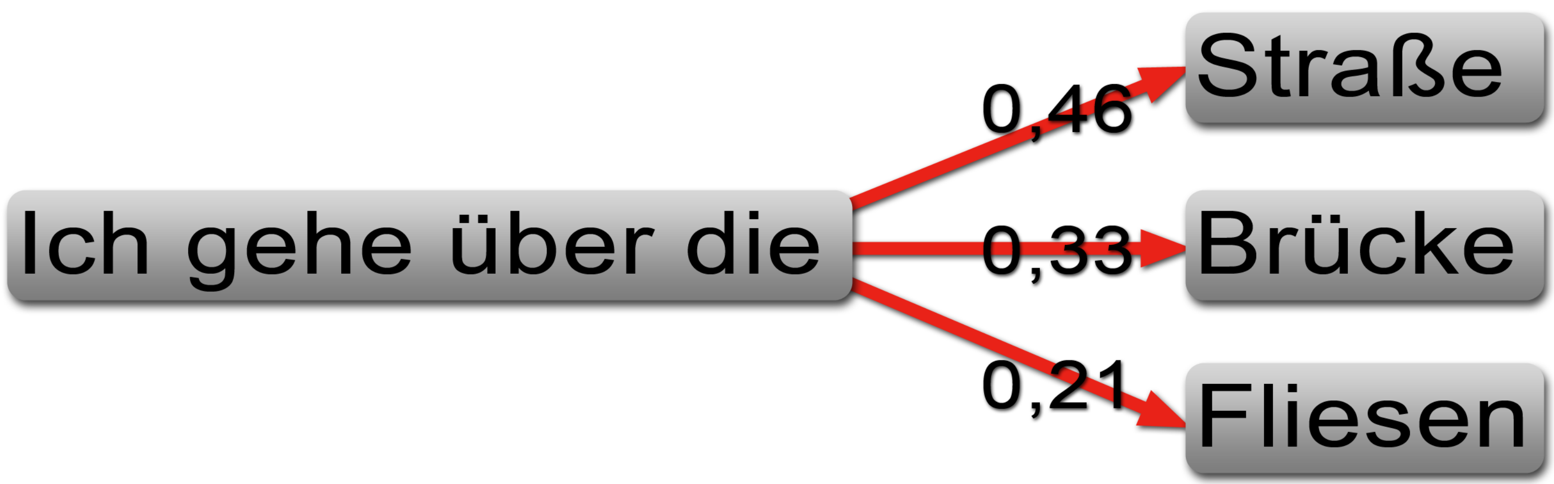
Outline of this talk

- Live Demonstration
- Discussion of the preliminary results (Motivation)
- Overview of the related work
- Introduction of our (yet informal) model
- An overview for the planned evaluation
- Outlook / Applications

The "mathematics" behind Language Models

w1 = Ich
w2 = gehe
w3 = über
w4 = die

search for: argmax{ P(w | w1, w2, w3, w4) }

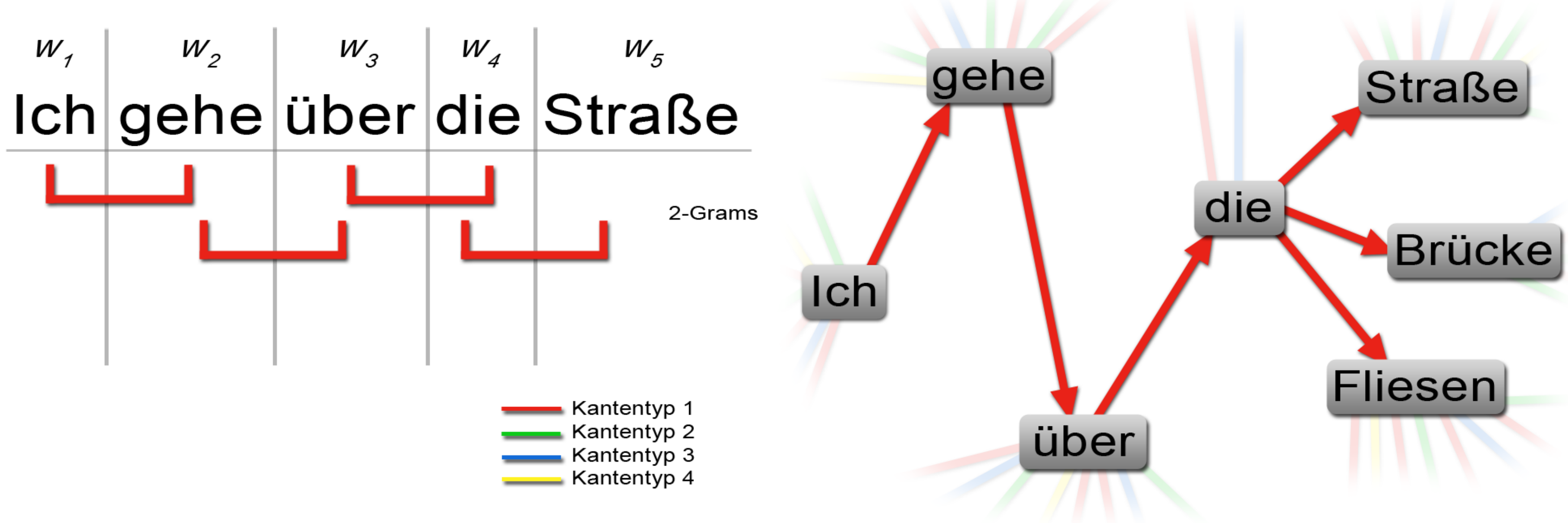


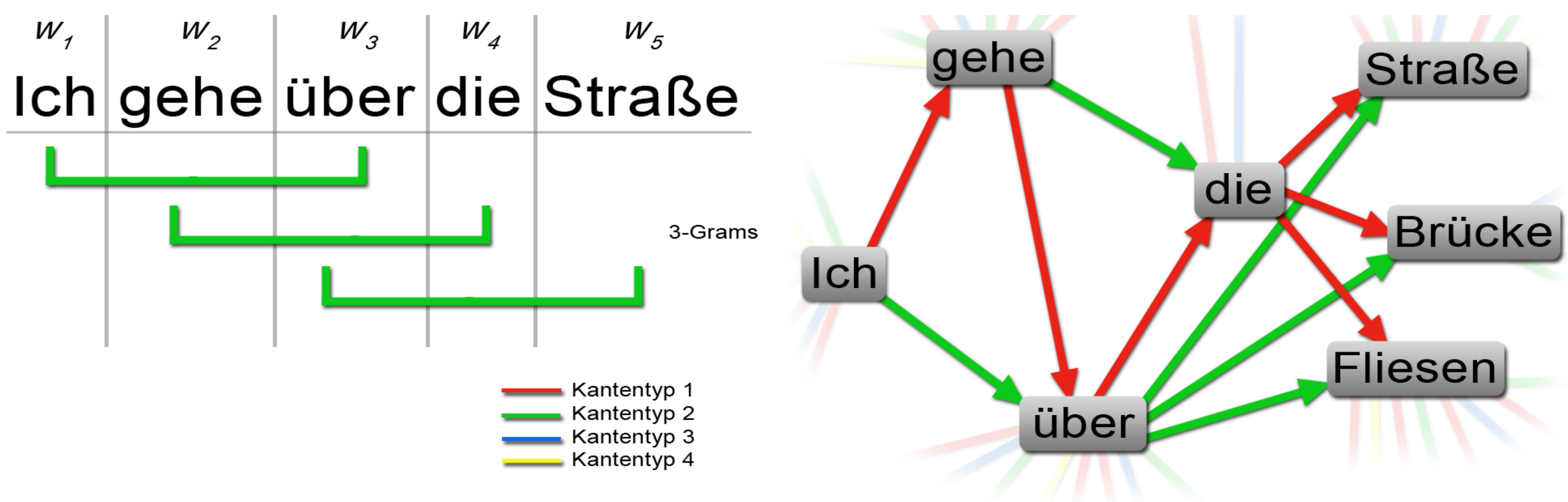
Look out! Data is very sparse and Zipf distribution

Rene Pickhardt
rpickhardt@uni-koblenz.de

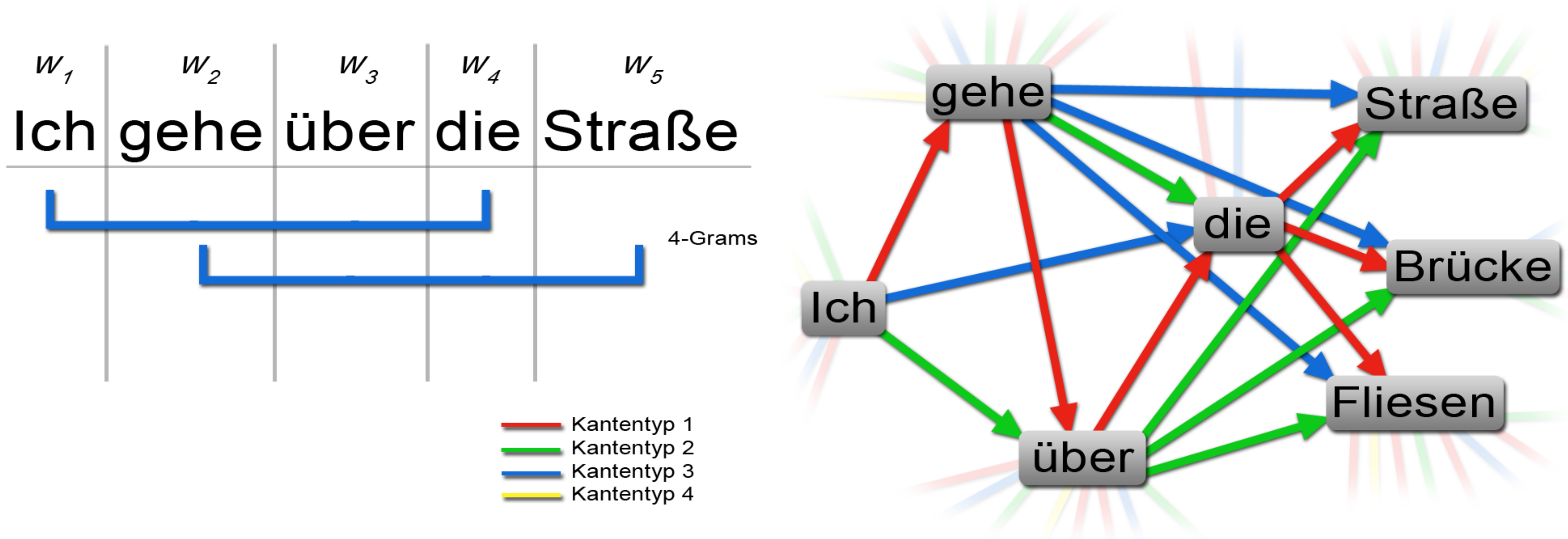
UNIVERSITÄT KOBLENZ · LANDAU

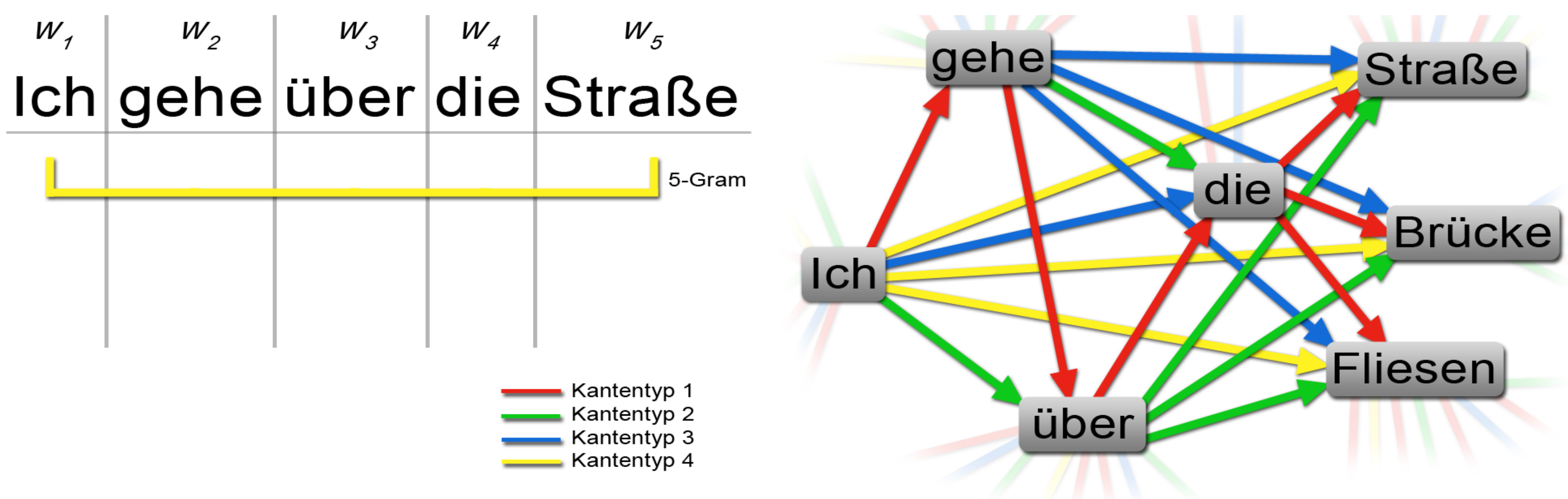
Start with 5 grams (available via Google Books)





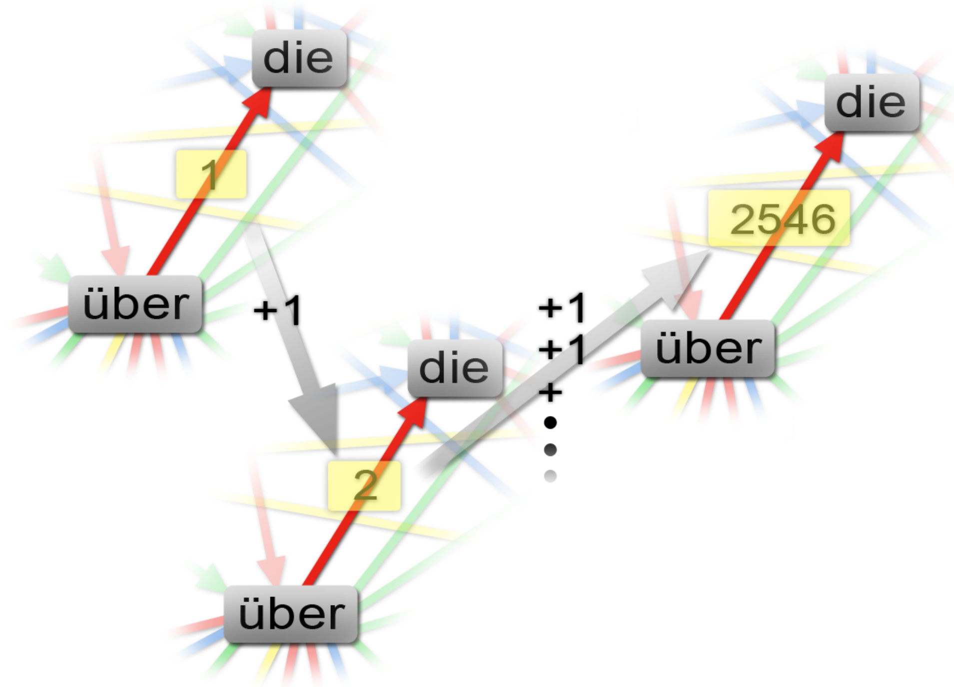
Start with 5 grams (available via Google Books)

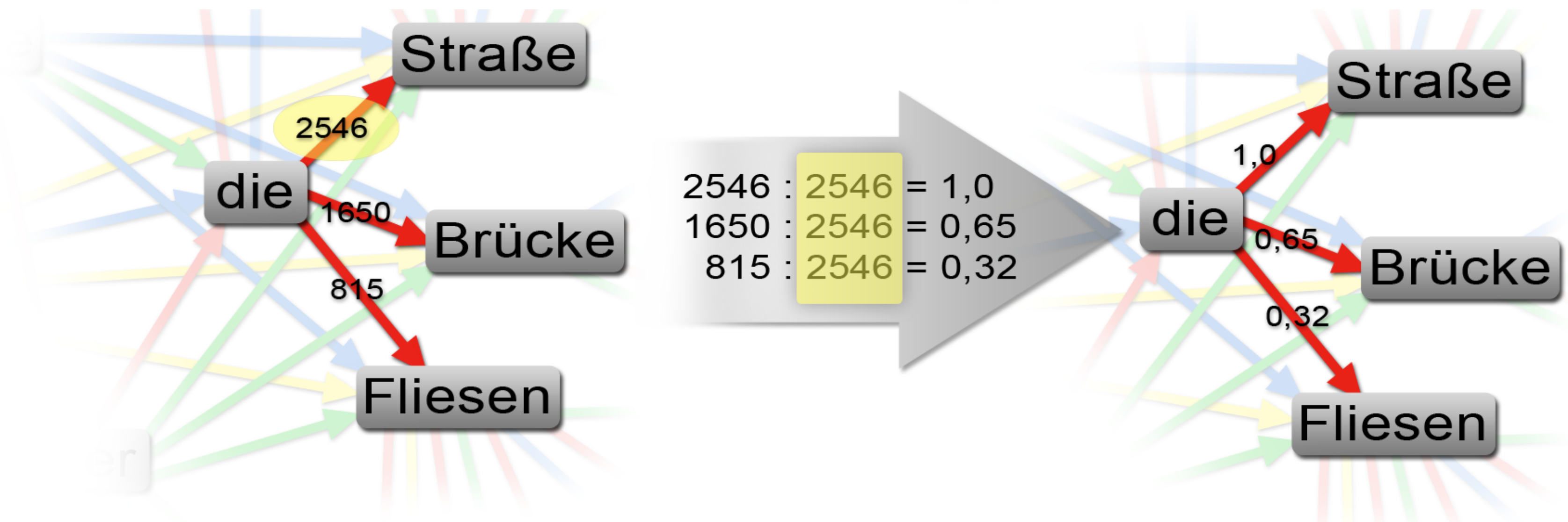




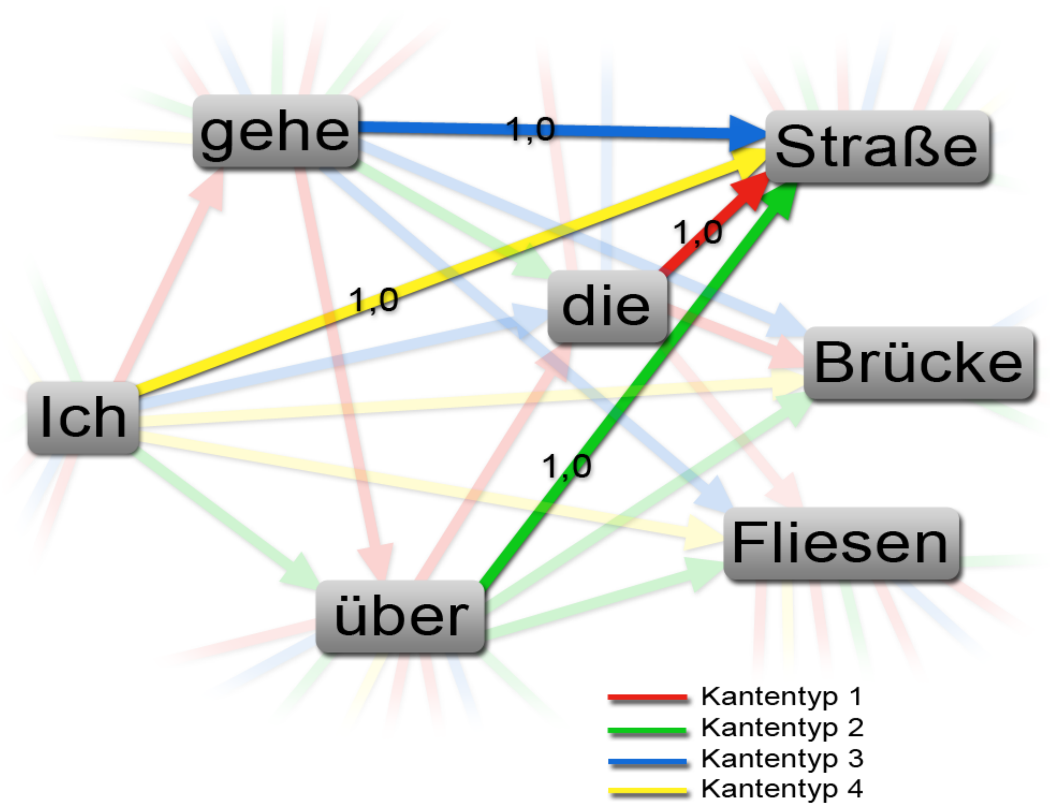
Normalized edge weights (for every edge type)

Ich gehe über die Straße. 1
+ +
Er fliegt über die Berge. 1
+ +
Ich gehe über die Brücke. 1
+ +
…arer Blick über die Stadt. 1
+ +
…mentation über die Geheim… 1
+ +
Ich gehe über die Fliesen. 1
+ +
…ie kommt über die Rücksch… 1
+ +
⋮ ⋮



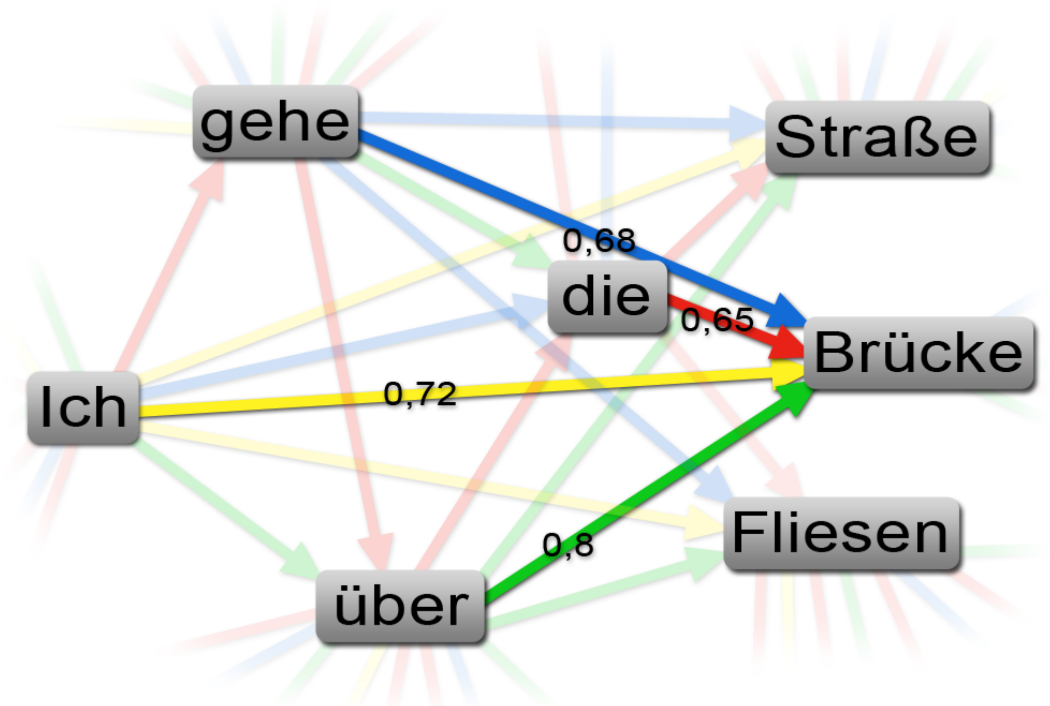


Retrieval algorithm



Straße: 1,0 + 1,0 + 1,0 + 1,0 = 4,0

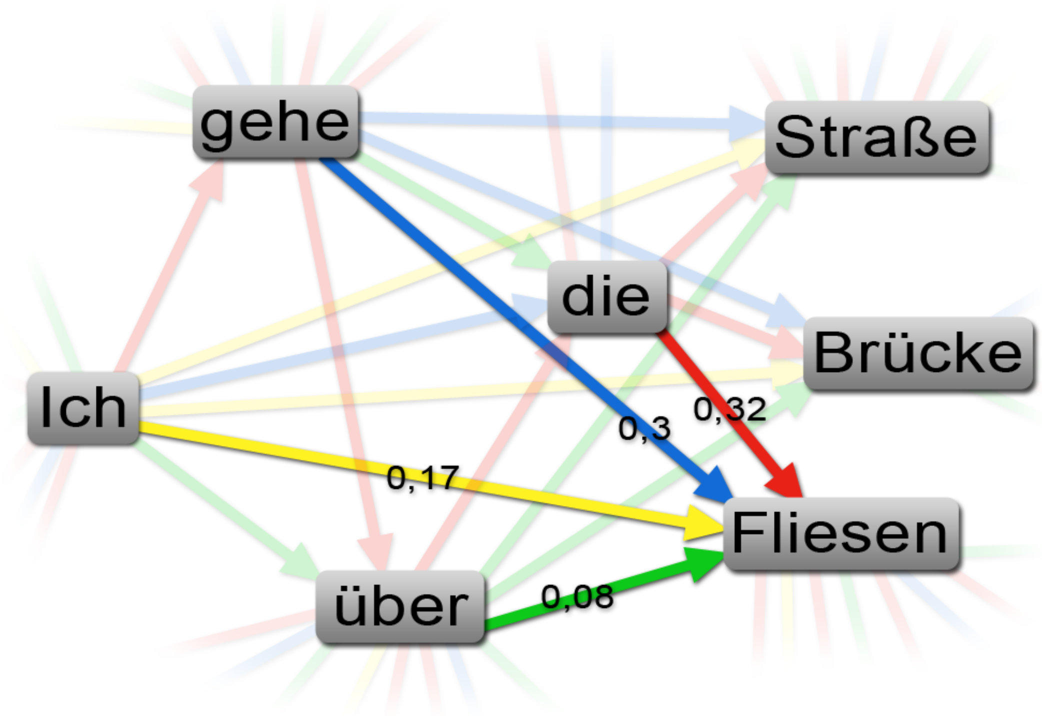
Retrieval algorithm



Straße: 1,0 + 1,0 + 1,0 + 1,0 = 4,0

Brücke: 0,72 + 0,68 + 0,8 + 0,65 = 2,85

Retrieval algorithm

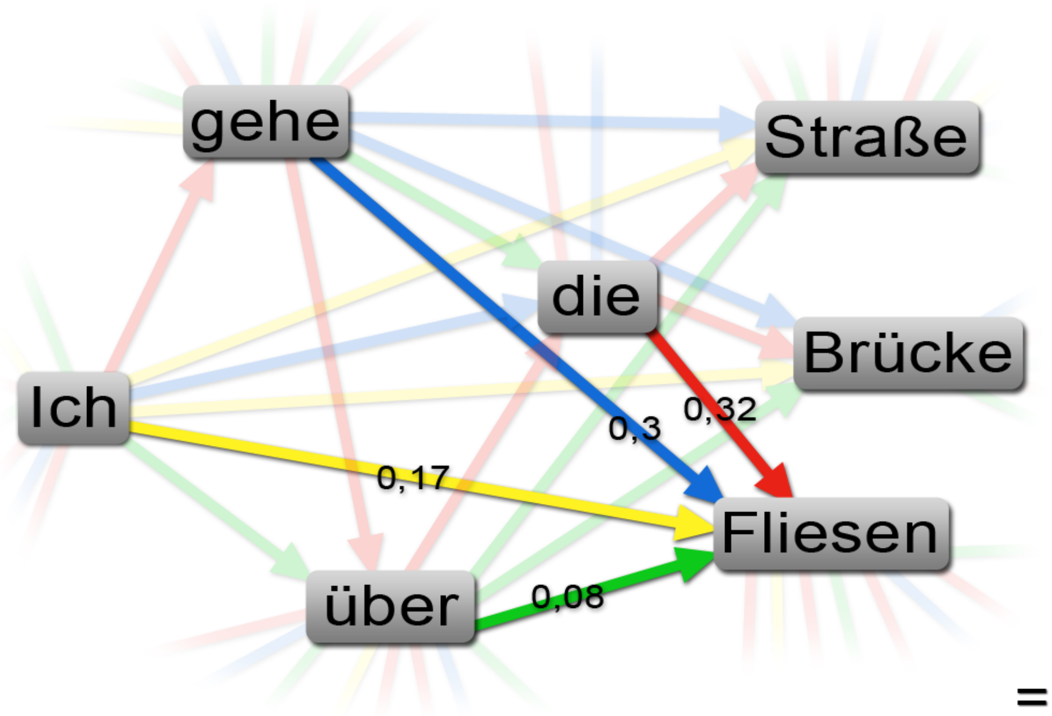


Straße: 1,0 + 1,0 + 1,0 + 1,0 = 4,0

Brücke: 0,72 + 0,68 + 0,8 + 0,65 = 2,85

Fliesen: 0,17 + 0,3 + 0,08 + 0,32 = 0,87

Retrieval algorithm



Straße: 1,0 + 1,0 + 1,0 + 1,0 = 4,0

Brücke: 0,72 + 0,68 + 0,8 + 0,65 = 2,85

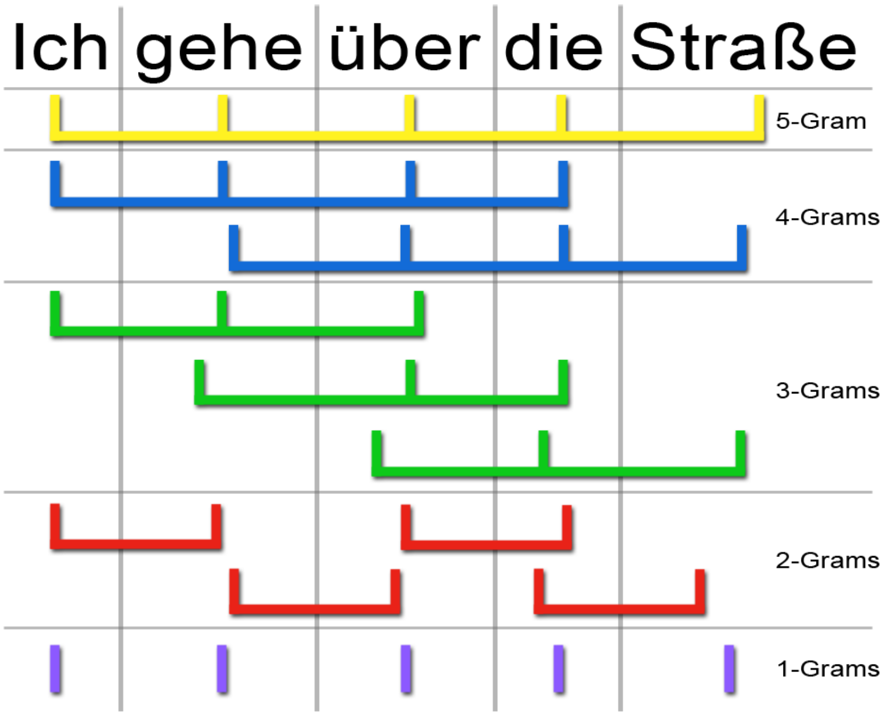
Fliesen: 0,17 + 0,3 + 0,08 + 0,32 = 0,87

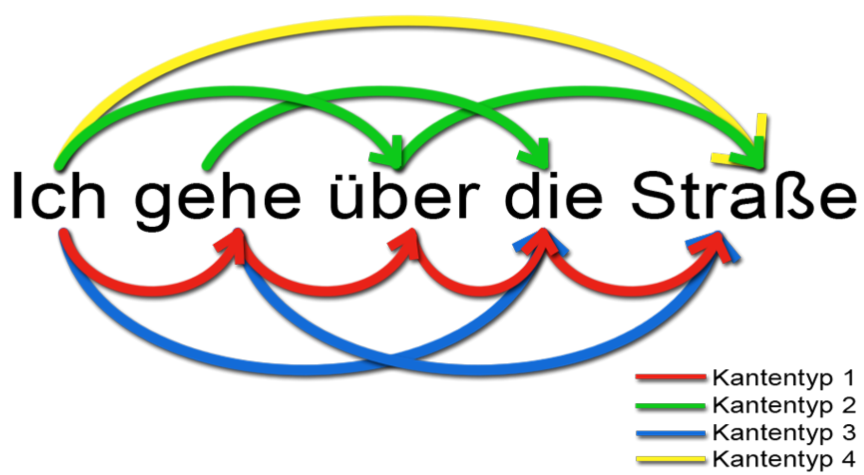
=> Straße > Brücke > Fliesen

Rene Pickhardt
rpickhardt@uni-koblenz.de

UNIVERSITÄT KOBLENZ · LANDAU

Start with 5 grams (available via Google Books)





- introduce n-distance
- n-gram leads to (n-1)-edge
- a 'generalized language' model with wild cards
  - 1-edge:
    - $P(w \mid w_4)$
  - 2-edge
    - $P(w \mid w_3, ?)$
  - 3-edge
    - $P(w \mid w_2, ?, ?)$
  - 4 edge
    - $P(w \mid w_1, ?, ?, ?)$

Retrieval of Typology is like Linear Interpolation

argmax $P_{Typology}(w \mid w_1, w_2, w_3, w_4)$ with

$$P_{Typology}(w \mid w_1, w_2, w_3, w_4) =$$

$$x_1 * P(w \mid w_4)$$

$$+ x_2 * P(w \mid w_3, ?)$$

$$+ x_3 * P(w \mid w_2, ?, ?)$$

$$+ x_4 * P(w \mid w_1, ?, ?, ?)$$

Retrieval of Typology is like Linear Interpolation

$$\textbf{argmax}\ P_{Typology}(w \mid w_1, w_2, w_3, w_4)\ \textbf{with}$$

$$P_{Typology}(w \mid w_1, w_2, w_3, w_4) =$$
$$x_1 * P(w \mid w_4)$$
$$+\ x_2 * P(w \mid w_3, ?)$$
$$+\ x_3 * P(w \mid w_2, ?, ?)$$
$$+\ x_4 * P(w \mid w_1, ?, ?, ?)$$

$$\textbf{argmax}\ P_{LM}(w \mid w_1, w_2, w_3, w_4)\ \textbf{with}$$

$$P_{LM}(w \mid w_1, w_2, w_3, w_4) =$$
$$y_1 * P(w \mid w_4)$$
$$+\ y_2 * P(w \mid w_3, w_4)$$
$$+\ y_3 * P(w \mid w_2, w_3, w_4)$$
$$+\ y_4 * P(w \mid w_1, w_2, w_3, w_4)$$

Efficient implementation is already achieved

WeST

- n-grams are preprocessed and stored in neo4j
  - takes quite some time (several hours)
  - ~80 GB n-grams compressed to ~1 GB neo4j DB
  - ~20 retrieval tasks per second
- Index using Suggest Tree's is created on top of neo4j db
  - takes ~2 minutes to build ~6 GB Index
  - ~14'000 retrieval tasks per second (on my notebook)
    - easy to distribute data structure of index
- Demo of Suggest Tree in different context available at:
  - http://gwt.metalcon.de/GWT-Modelling/#AutoCompletionTest

Outline of this talk

- Live Demonstration
- Discussion of the preliminary results (Motivation)
- Overview of the related work
- Introduction of our (yet informal) model
- An overview for the planned evaluation
- Outlook / Applications

We want to understand / evaluate...

- does graph based n-distance depend on
  - the used language
  - special domain corpora
  - data sparsity
  - the lenght of n-distance (how many edges do we need?)
    - Entropy of n-distance
- behavior with respect to base lines
  - Language Models
    - Linear interpolation
    - maximum Likelihood Estimation
  - modern Baselines from related work (yet undecided)

Quality measures

- Precision / Accuracy
- Keystroke savings
- MRR (mean reciprocal rank)
- maybe a user study / user experiment
- any other?

rpickhardt@uni-koblenz.de
UNIVERSITÄT KOBLENZ · LANDAU

Roadmap (interesting data sets)

- **Wikipedia**
  - general purpose
  - multilingual
  - learning / testing
- **Google ngrams**
  - general purpose
  - multilingual
  - learning
- **Reuters**
  - special purpose (news domain)
  - multi lingual
  - testing
- **EU Protocols (JRC Acquis 2012)**
  - special purpose (politics)
  - multi lingual
  - testing

Outline of this talk

- Live Demonstration
- Discussion of the preliminary results (Motivation)
- Overview of the related work
- Introduction of our (yet informal) model
- An overview for the planned evaluation
- Outlook / Applications

This model can be applied to the fields of

- speech recognition
- grammar correction
- Machine translation
- improve HCI
- personalized text recommendations
    - possibly through sparse data requirements

Thank you for your attention

More information + Slides on:
http://www.typology.de

http://www.rene-pickhardt.de/tag/typology

Android app available at:
http://www.typology.de/android-app

Special thanks to Till and Paul for implementing and testing my initial idea for the sake of Jugend Forscht.