Generating Acrostics via Paraphrasing and Heuristic Search

Bruno Soares Fillmann
Fernando Bombardelli da Silva
Jürgen Bauer
William Bombardelli da Silva

Technische Universität Berlin
Datenbanksysteme und Informationsmanagement
DBPRO – Database Projects (WS 2014/2015)

12.01.2015

Organization

The Problem

- Given a text T and an acrostic X, find a paraphrased version of T that encodes X in the first letters of each line.

Knuth ist der Sohn eines Lehrers. Er besuchte die Milwaukee Lutheran
High School und begann sein Physikstudium am California Institute of
Technology im September 1956. Aus zweierlei Gründen schlug er
tatsächlich seinem zweiten Studienjahr jedoch den Weg zur Mathematik
ein: Zum einen löste er ein Problem eines seiner
Mathematikprofessoren, was ihm eine 1,0 als Note einbrachte, zum
anderen fand er wenig Gefallen an den physikalischen Praktika. Er
erhielt einen Bachelor- und einen Master-Abschluss 1960 an der Case
Western Reserve University.

Result Text:

Knuth ist der Sohn eines Lehrers. Er besuchte die Milwaukee Luthera- *//Wrong Hyphenation*
n High School und begann sein Physikstudium am California Instit- *//Line Break +*
ute of Technology im September 1956. Aus zweierlei Gründen schlug er *//Wrong Hyphenation*
tatsächlich seinem zweiten Studienjahr jedoch den Weg zur Mat- *//Wrong Hyphenation*
hematik ein: Zum einen löste er ein Problem eines seiner
Mathematikprofessoren, was ihm eine 1,0 als Note einbrachte, zum
anderen fand er wenig Gefallen an den physikalischen Praktika. Er
erhielt einen Bachelor- und einen Master-Abschluss 1960 an der Case
Western Reserve University.

The Project

- **Main goal of the project:** Implement the paper for the German language.
- **Main idea of the algorithm:**
  - Modeled as a search problem in a tree.
  - The vertices are states (texts) and the edges are operations over states.
  - Artificial intelligence is applied for the search strategy (A* Algorithm).

Last State of Work

- Line Break and Wrong Hyphenation working properly
- Search algorithm working properly
- Word Insertion working partially, but Deletion not working

New Developed Operations

- Word Insertion Or Deletion (Enhancements)
- Synonym
- Hyphenation
- Spelling

Synonym

- Free dictionary of synonyms is used, namely, Open Thesaurus (www.openthesaurus.de)
- We use **Redis** as database server (NoSQL)
- Redis stores data as key-value pairs
- Our base is structured in a way that, every word in the thesaurus is a key that points to a set of synonyms
- The thesaurus text file is imported into the database by a Python script
- The database is then consulted by the application during its execution

Hyphenation

- Line Length constraints $l_{min} = 50$ and $l_{max} = 70$.
- For hyphenation a re-implementation of Knuth's hyphenation algorithm in TeX is used (TEXHyphenator-J by David Tolpin)
- A word can be hyphenated if: The part of the current line **from** the beginning **to** the hyphen has a length of at least $l_{min} = 50$.
- The text **after** the hyphen has to be aligned again to fulfill the line length constraints.
- A greedy word wrap algorithm is applied.
- Don't allow words of length > 20 in the start text.

Hyphenation

- **Example:**
- 38 hyphenations!

ORIGINAL TEXT:
Knuth ist der Sohn eines Lehrers. Er besuchte die Milwaukee Lutheran High School und begann sein Physikstudium .....

0.
Knuth ist der Sohn eines Lehrers. Er besuchte die Mil-
waukee Lutheran High School und begann sein Physikstudium .....

1.
Knuth ist der Sohn eines Lehrers. Er besuchte die Milwau-
kee Lutheran High School und begann sein Physikstudium .....

2.
Knuth ist der Sohn eines Lehrers. Er besuchte die Milwaukee Lu-
theran High School und begann sein Physikstudium .....

Spelling

- This operation adds wrong spellings of words to add letter variety
- To do this it changes letters such as 'ä', 'ö' and 'ü' to respectively ae oe ue.
- The german letter 'ß' can also be changed to 'ss'
- Only one word is changed per new text created by this operation

Results – Spelling

Original Text:

```
Die etymologischen Vorformen von "deutsch" bedeuteten ursprünglich
"zum Volk gehörig", wobei das Adjektiv zunächst die Dialekte des
kontinental-westgermanischen Dialektkontinuums bezeichnete.  Die
Bezeichnung Deutschland wird seit dem 15.  Jahrhundert verwendet, ist
in einzelnen Schriftstücken aber schon früher bezeugt.
```

Result Text:

```
Die etymologischen Vorformen von "deutsch" bedeuteten urspru-
englich "zum Volk gehörig", wobei das Adjektiv zunächst die Dialekte de-
s kontinental-westgermanischen Dialektkontinuums bezeichnete.  Die
Bezeichnung Deutschland wird seit dem 15.  Jahrhundert verwendet, ist
in einzelnen Schriftstücken aber schon früher bezeugt.
```

Results – Insertion Of New Words

Im Englischen bedeutet blau sein so etwas wie traurig oder deprimiert sein, weltbekannt ist die Musikrichtung Blues. Das ist traurig aber. Im Deutschen hat blau sein allerdings eine ganz andere Bedeutung. Wenn jemand eine feucht-fröhliche Party feiert (also eine, auf der viel Alkohol getrunken wird), dann ist er hinterher wahrscheinlich blau. Blau sein bedeutet im Deutschen nämlich betrunken sein.

Result Text:

Im Englischen bedeutet blau sein so etwas wie traurig oder *//Line Break*
deprimiert sein, weltbekannt ist die Musikrichtung Blues. Das ist
echt traurig aber. Im Deutschen hat blau sein allerdings *//Word Insertion*
eine ganz andere Bedeutung. Wenn jemand eine feucht-fröhliche Party *//+ Line Break*
feiert (also eine, auf der viel Alkohol getrunken wird), dann ist er
hinterher wahrscheinlich blau. Blau sein bedeutet im Deutschen nämlich
betrunken sein.

Results – Deletion Of Words

In diesen Auflagen könnte denn auch der Hebel liegen, um Griechenland weiteres Geld zu verweigern. Denn ein Stopp der Zinszahlungen dürfte als Verstoß gegen das Rettungsprogramm gewertet werden. Denn für alle Geschäftbanken in der Eurozone gilt: Geld von der EZB gibt es nur mäßig, wenn die Finanzinstitute eine bestimmte Menge an Wertpapieren, zum Beispiel Staatsanleihen, als Pfand bei der EZB hinterlegen. In diesem Fall könnte die EZB ihre Ansprüche an die Sicherheiten wieder erhöhen.

Result Text:

In diesen Auflagen könnte denn auch der Hebel liegen, um Griechenland weiteres Geld zu verweigern. Denn ein Stopp der Zinszahlungen dürfte als Verstoß gegen das Rettungsprogramm gewertet werden. Denn für alle in der Eurozone gilt: Geld von der EZB gibt es nur Mäßig, wenn die Finanzinstitute eine bestimmte Menge an Wertpapieren, zum Beispiel Staatsanleihen, als Pfand bei der EZB hinterlegen. In diesem Fall könnte die EZB ihre Ansprüche an die Sicherheiten wieder erhöhen.

//Wrong Hyphenation
//Wrong Hyphenation
//Word Deletion

Results – Synonym

Deutsches Leben der Gegenwart -- dem feindlichen Blick, der **nur** seine Oberfläche streift, möchte scheinen, daß die Gegenwart **wenig** vom deutschen Leben, mehr vom deutschen Sterben zu melden hätte. **Aber** der nachdenkliche Betrachter weiß, daß die größten geistigen Epochen Deutschlands über seinen politischen Niederlagen wuchsen, daß **gerade** die Zeiten nach dem Dreißigjährigen Krieg, nach dem Zusammenbruch von Jena zu den schöpferischen des deutschen Lebens gehören.

Deutsches Leben der Gegenwart -- dem feindlichen Blick, der **gerade**	**// Synonym**
eben seine Oberfläche streift, möchte scheinen, daß die Gegenwart	**// Line break**
unbedeutend vom deutschen Leben, mehr vom deutschen Sterben zu melden hätte.	**// Synonym**
trotzdem der nachdenkliche Betrachter weiß, daß die größten geistigen	**// Synonym**
Epochen Deutschlands über seinen politischen Niederlagen wuchsen, daß	**// Line break**
nachgerade die Zeiten nach dem Dreißigjährigen Krieg, nach dem	**// Synonym**
Zusammenbruch von Jena zu den schöpferischen des deutschen Lebens gehören.	

Results – Hyphenation

- 1 min 22 sec, 17517 nodes, Ops: Hyph,WrHyph,LB

ORIGINAL TEXT:
Friedrich wurde im Berliner Stadtschloss geboren und war der älteste überlebende Sohn von insgesamt 14 Kindern König Friedrich Wilhelms I. und dessen Gattin Sophie Dorothea, der ersten hochadligen Königin des Fürstentums zu Hannover. (WrongHyphenation)

Friedrich wurde im Berliner Stadtschloss geboren und wa-
r der älteste überlebende Sohn von insgesamt 14 Kindern König Friedrich Wilhelms I. und dessen Gattin Sophie Dorothea, der ersten hochadligen Königin des Fürstentums zu Hannover. (WrongHyphenation)

Friedrich wurde im Berliner Stadtschloss geboren und wa-
r der älteste überlebende Sohn von insgesamt 14 Kindern König Fr-
iedrich Wilhelms I. und dessen Gattin Sophie Dorothea, der ersten hochadligen Königin des Fürstentums zu Hannover. (Hyphenation)

Examples

- 1 min 22 sec, 17517 nodes, Ops: Hyph,WrHyph,LB

Friedrich wurde im Berliner Stadtschloss geboren und wa-
r der älteste überlebende Sohn von insgesamt 14 Kindern König Fr-
iedrich Wilhelms I. und dessen Gattin Sophie Dorothea, der ersten
hochadligen Königin des Fürstentums zu Hannover. (Hyphenation)

Friedrich wurde im Berliner Stadtschloss geboren und wa-
r der älteste überlebende Sohn von insgesamt 14 Kindern König Fr-
iedrich Wilhelms I. und dessen Gattin Sophie Doro-
thea, der ersten hochadligen Königin des Fürstentums zu Hannover. (LineBreak)

Friedrich wurde im Berliner Stadtschloss geboren und wa-
r der älteste überlebende Sohn von insgesamt 14 Kindern König Fr-
iedrich Wilhelms I. und dessen Gattin Sophie Doro-
thea, der ersten hochadligen Königin des Fürstentums
zu Hannover.

Conclusion

- **Operations:** Line Break, Hyphenation, Wrong Hyphenation, Synonym, Word Insertion or Deletion and Spelling.
- **What the algorithm does:**
  - Build acrostics with short words (until 6 letters) in reasonable time
  - Build acrostics with not so good quality
  - Success depends highly on certain conditions of the text (e.g. most of the letters of the acrostic must be in)
- **Difficulties:**
  - The time consumed by the requests to the NGram web service
  - The quality of the NGram database
  - The use of NGram is sometimes too restrictive
  - The choice of the quality of each operation is decisive in the result

References


Benno Stein, Matthias Hagen, and Christof Bräutigam. *Generating Acrostics via Paraphrasing and Heuristic Search*.
In Junichi Tsujii and Jan Hajic, editors, 25th International Conference on Computational Linguistics (COLING 14), pages 2018-2029, August 2014. Association for Computational Linguistics.


Spiegel Online. *Griechenland und die Euro-Zone: Der fast unmögliche Rausschmiss*.
http://www.spiegel.de/wirtschaft/soziales/griechenland-euro-austritt-waere-moeglich-aber-kompliziert-a-1011361.html. Am 07. Januar 2015


Hilko. *Blau ist keine traurige Farbe*.
https://deutschlich.wordpress.com/2013/03/12/blau-ist-keine-traurige-farbe. Am 07. Januar 2015


Redis. http://redis.io/. Am 07. Januar 2015


Synonym - OpenThesaurus - Deutscher Thesaurus.
https://www.openthesaurus.de/. Am 07. Januar 2015

Questions?