Horváth & Partners GmbH
Technical University of Munich

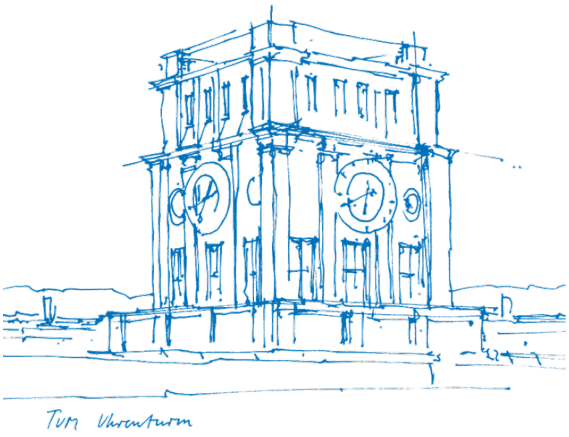
TUM Data Innovation Lab

NLP for intelligent data mining

Simon Lohrmann, Vaibhav Jain, Niklas Lüdtke, Esmée Oosterlaar

Horváth & Partners GmbH
Technical University of Munich

Summersemester 2022



Introduction

Goal

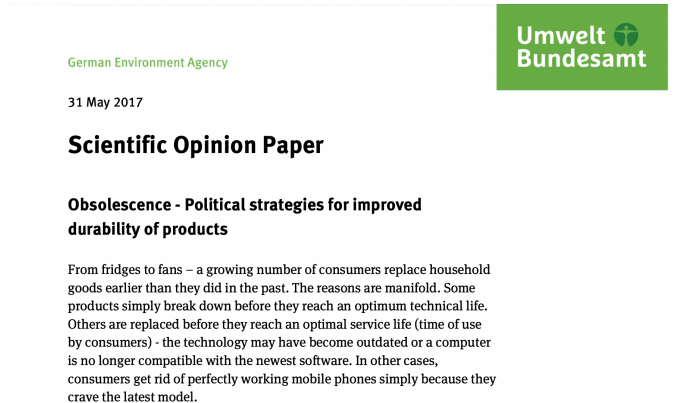
Investigate relationships between entities extracted from text documents and build a knowledge graph

Focus & Contribution

1. Investigate process: PDF $\rightarrow$ knowledge graph
2. Merge models & algorithms into end-to-end pipeline

Data from German Environment Agency

- Scientific papers in English and German language



German Environment Agency

31 May 2017

Scientific Opinion Paper

Obsolescence - Political strategies for improved durability of products

From fridges to fans – a growing number of consumers replace household goods earlier than they did in the past. The reasons are manifold. Some products simply break down before they reach an optimum technical life. Others are replaced before they reach an optimal service life (time of use by consumers) - the technology may have become outdated or a computer is no longer compatible with the newest software. In other cases, consumers get rid of perfectly working mobile phones simply because they crave the latest model.

Text documents contain entities and relationships

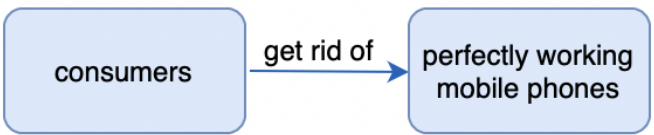
- Entities are real world objects such as persons, locations etc.

> In other cases, consumers get rid of perfectly working mobile phones simply because they crave the latest model.

- Some entity pairs have a relation

> In other cases, consumers get rid of perfectly working mobile phones simply because they crave the latest model.

Graph visualizes entities and their relationships



Project Pipeline


Abstract Extraction
Coreference Resolution
Entity Extraction
Relationship Classification
Knowledge Graph

Abstract contains most important information

German Environment Agency

31 May 2017

Scientific Opinion Paper

Obsolescence - Political strategies for improved durability of products

From fridges to fans – a growing number of consumers replace household goods earlier than they did in the past. The reasons are manifold. Some products simply break down before they reach an optimum technical life. Others are replaced before they reach an optimal service life (time of use by consumers) - the technology may have become outdated or a computer is no longer compatible with the newest software. In other cases, consumers get rid of perfectly working mobile phones simply because they crave the latest model.

Abstract extraction focuses on font size

Problem

PDF’s don’t store high-level features of documents

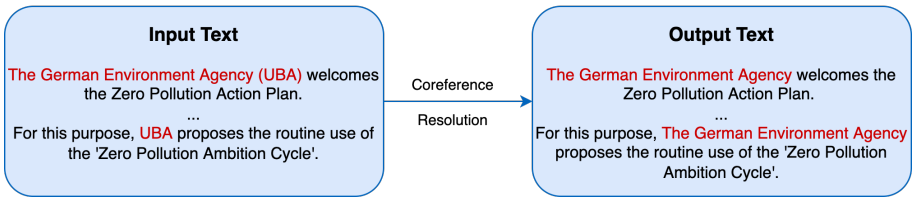
Solution

1. Get font size of paragraph by heuristics
2. Extract all text with the same properties, mainly size
3. Search for predefined keywords like "Abstract"
4. Set a minimum character count to avoid false positives


Abstract Extraction
Coreference Resolution
Entity Extraction
Relationship Classification
Knowledge Graph

Coreference resolution

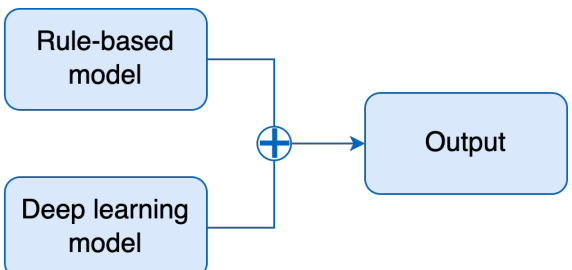
- Needed if two or more expressions refer to the same entity
- Duplicate edges in knowledge graph are avoided
- Long range dependencies can be handled
- AllenNLP coreference predictor




Abstract Extraction
Coreference Resolution
Entity Extraction
Relationship Classification
Knowledge Graph

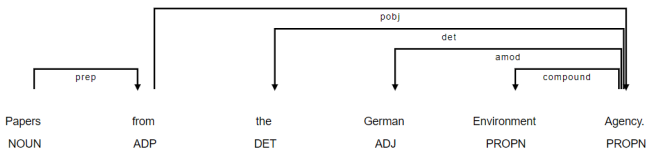
Entity extraction

- First major step of the pipeline: find entities where a relation is likely to occur
- Entities are real world objects such as persons, locations etc.
- String is split in several strings each consisting of a sentence



Rule-based approach

- Predefined rules to identify entities that follow a certain pattern
- Idea: subject and object of a sentence are most likely to be entities
- Extract subjects, objects along with its modifiers, compound words and punctuation marks
- High precision (measure of quality) as words extracted are most likely to be entities
- Moderate recall (measure of quantity) as not all entities might be subject or object



Pre-trained named-entity recognition

- Most relationships between organizations, persons etc.
  - Use named-entity recognition (NER)
- NER modeled as Deep Neural Network
  - CNN or Transformer architecture
- Pre-trained models from spaCy
- Transformer model for English text documents
- CNN model for German text documents
- Omit cardinal numbers and dates

Example of spaCy transformer model

The European Green Deal´s zero pollution vision to 2050 and the Chemicals Strategy for Sustainability have raised the level of ambition of the European Union´s policies on chemicals.
Against this background, the targeted revision of the REACH regulation offers a unique and timely opportunity to further strengthen the legislative text.
This paper focuses on the revision of the REACH authorisation and restriction system from an environmental perspective.
the REACH authorisation and restriction system are central for the regulation of the continued use of the most hazardous substances, as well as the manufacture, placing on the market or use of substances when there is an unacceptable risk.
While This paper deliberately does not take a position visàvis the policy options discussed by others, our recommendations mostly support what has been proposed by the European Commission so far.
Based on an analysis of the strengths and weaknesses of the current REACH authorisation and restriction system, This paper recommends six objectives and ten buildingblocks as well as procedural steps for the REACH authorisation and restriction system.
Buildingblocks offer a flexible approach that can be adapted easily to different policy options.

Combining models improves entity extraction

- Extracted entities for the deep learning approach:

> The European Green Deal's zero pollution vision to 2050 and the Chemicals Strategy for Sustainability have raised the level of ambition of the European Union's policies on chemicals.

- Extracted entities for the rule-based approach:

> The European Green Deal's zero pollution vision to 2050 and the Chemicals Strategy for Sustainability have raised the level of ambition of the European Union's policies on chemicals.

- All extracted entities combined from both models:

> The European Green Deal's zero pollution vision to 2050 and the Chemicals Strategy for Sustainability have raised the level of ambition of the European Union's policies on chemicals.

Finalizing entity extraction by combining models

The European Green Deal´s zero pollution vision to 2050 and the Chemicals Strategy for Sustainability have raised the level of ambition of the European Union´s policies on chemicals.
Against this background, the targeted revision of the REACH regulation offers a unique and timely opportunity to further strengthen the legislative text.
This paper focuses on the revision of the REACH authorisation and restriction system from an environmental perspective.
the REACH authorisation and restriction system are central for the regulation of the continued use of the most hazardous substances, as well as the manufacture, placing on the market or use of substances when there is an unacceptable risk.
While This paper deliberately does not take a position visàvis the policy options discussed by others, our recommendations mostly support what has been proposed by the European Commission so far.
Based on an analysis of the strengths and weaknesses of the current REACH authorisation and restriction system, This paper recommends six objectives and ten buildingblocks as well as procedural steps for the REACH authorisation and restriction system.
Buildingblocks offer a flexible approach that can be adapted easily to different policy options.

Preparation for relationship extraction

- Mark the entities for relationship extraction and knowledge graph generation
- Consider sentence multiple times, if more than two entities in one sentence

[E1]The European Green Deal's[/E1] zero pollution vision to 2050 and [E2]the Chemicals Strategy for Sustainability[/E2] have raised the level of ambition of the European Union's policies on chemicals.

[E1]The European Green Deal's[/E1] zero pollution vision to 2050 and the Chemicals Strategy for Sustainability have raised the level of [E2]ambition[/E2] of the European Union's policies on chemicals.

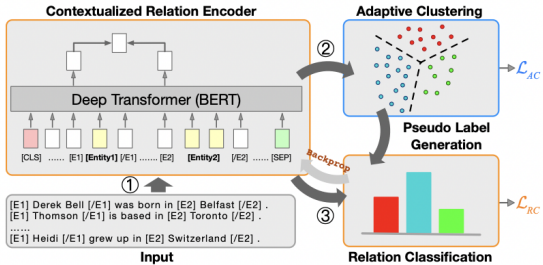
The European Green Deal's zero pollution vision to 2050 and [E1]the Chemicals Strategy for Sustainability[/E1] have raised the level of [E2]ambition[/E2] of the European Union's policies on chemicals.

...


Abstract Extraction
Coreference Resolution
Entity Extraction
Relationship Classification
Knowledge Graph

SelfORE: a self-supervised model

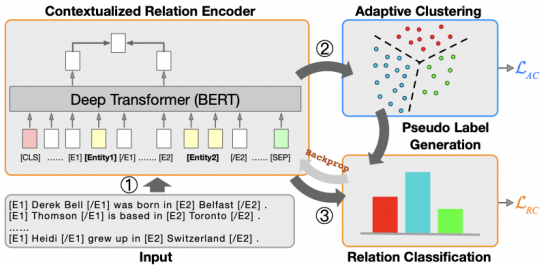
- Self-supervised Relational Feature Learning for Open Relation Extraction



SelfORE model [Hu+20]

SelfORE: a self-supervised model

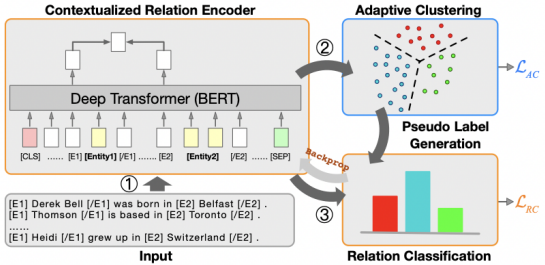
- Self-supervised Relational Feature Learning for Open Relation Extraction
- No relations specified in advance



SelfORE model [Hu+20]

SelfORE: a self-supervised model

- Self-supervised Relational Feature Learning for Open Relation Extraction
- No relations specified in advance
- Contextualized Relation Encoder (RE), Adaptive Clustering (AC), Relation Classification (RC)

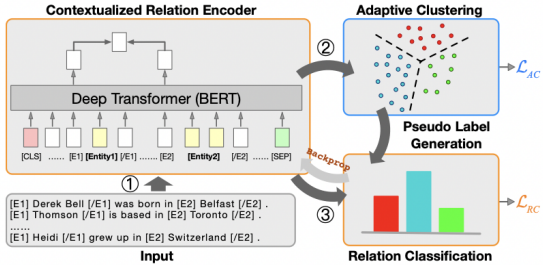


SelfORE model [Hu+20]

SelfORE consists of three main parts

- Cont. Relation Encoder

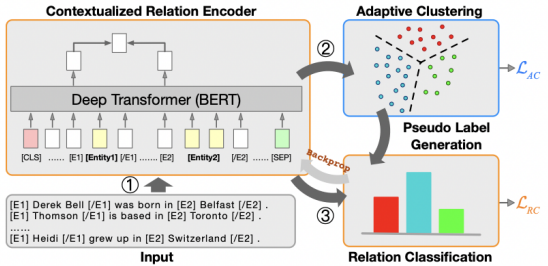
  Use pre-trained transformer model from Huggingface, e.g. Bert



SelfORE model [Hu+20]

SelfORE consists of three main parts

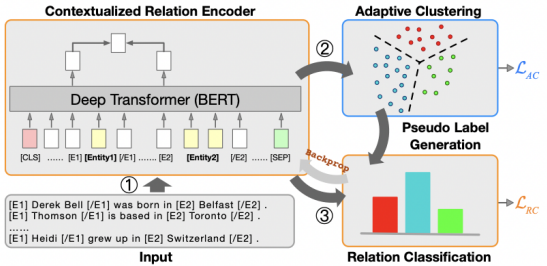
- Cont. Relation Encoder
  Use pre-trained transformer model from Huggingface, e.g. Bert
- Adaptive Clustering
  - Encode hidden state of RE
  - Learn enconder layers by using cluster centroids from KMeans
  - Soft-assign each sample to the centroids
  - Stop training when there is no change in pseudo labels



SelfORE model [Hu+20]

SelfORE consists of three main parts

- Cont. Relation Encoder
  Use pre-trained transformer model from Huggingface, e.g. Bert
- Adaptive Clustering
  - Encode hidden state of RE
  - Learn enconder layers by using cluster centroids from KMeans
  - Soft-assign each sample to the centroids
  - Stop training when there is no change in pseudo labels
- Relation Classification
  MLP trained from scratch



SelfORE model [Hu+20]

How can we test the model?

How can we test the model?

Unlabelled data:

Labelled data:

How can we test the model?

Unlabelled data:

- Measure prediction power of classification for pseudo labels
- Cluster distribution

Labelled data:

How can we test the model?

Unlabelled data:

- Measure prediction power of classification for pseudo labels
- Cluster distribution

Labelled data:

- Assign majority label to each cluster
- Compare predicted labels to ground-truth labels, using: B^3, V-measure, ARI
- Use most frequent n-gram in the text between entities as relation

Results for SelfORE

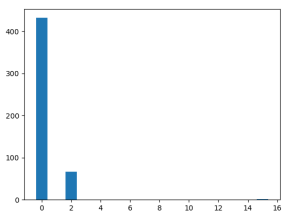
Unlabelled data: (Papers)

Labelled data: (T-REx SPO)

Results for SelfORE

Unlabelled data: (Papers)

- Prediction acc. of classification for pseudo labels $> 95\%$
- Cluster distribution by index:

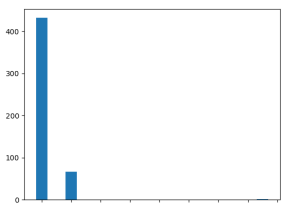


Labelled data: (T-REx SPO)

Results for SelfORE

Unlabelled data: (Papers)

- Prediction acc. of classification for pseudo labels $> 95\%$
- Cluster distribution by index:



Labelled data: (T-REx SPO)

	Our	[Hu+20]
B^3	8.0	41.0
V-Meas	5.2	41.4
ARI	0.8	33.7

Note: Our model was evaluated on a small subset of T-REx SPO!

Creating labels with heuristics

N-gram

Root word

Creating labels with heuristics

N-gram

- Most frequent n-gram (window of n contiguous words from text) between entities as relationship
- Results: for training on T-REx subset, 3-grams do not perform well due to lack of available data to generate 3-grams

Root word

Extracted surface-form	Golden surface-form
are close to	shares border with
the state of	country
capital city	capital
son of	child
member of	member of

Figure 1 Example from [Hu+20]

Creating labels with heuristics

N-gram

- Most frequent n-gram (window of n contiguous words from text) between entities as relationship
- Results: for training on T-REx subset, 3-grams do not perform well due to lack of available data to generate 3-grams

Extracted surface-form	Golden surface-form
are close to	shares border with
the state of	country
capital city	capital
son of	child
member of	member of

Figure 1 Example from [Hu+20]

Root word

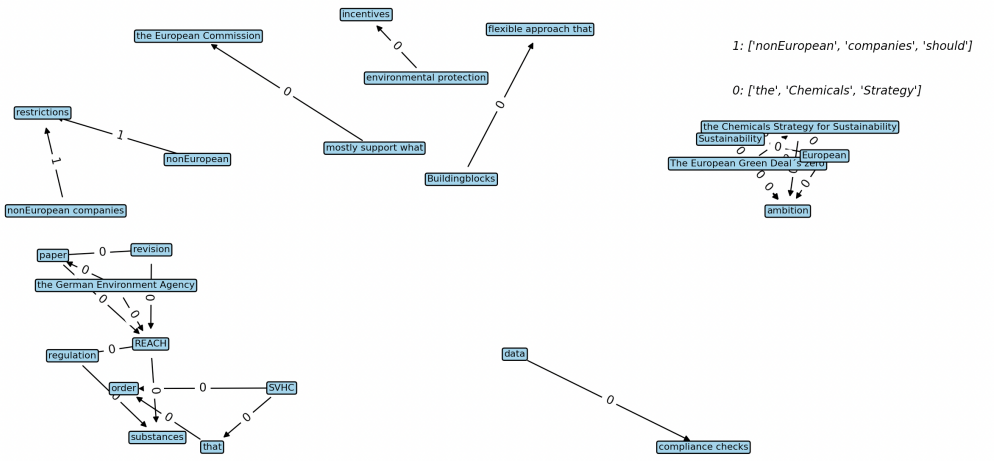
- Root word (main verb) encodes information of relation
- More consistent results with small dataset
- But: more data $\rightarrow$ better performance


Abstract Extraction
Coreference Resolution
Entity Extraction
Relationship Classification
Knowledge Graph

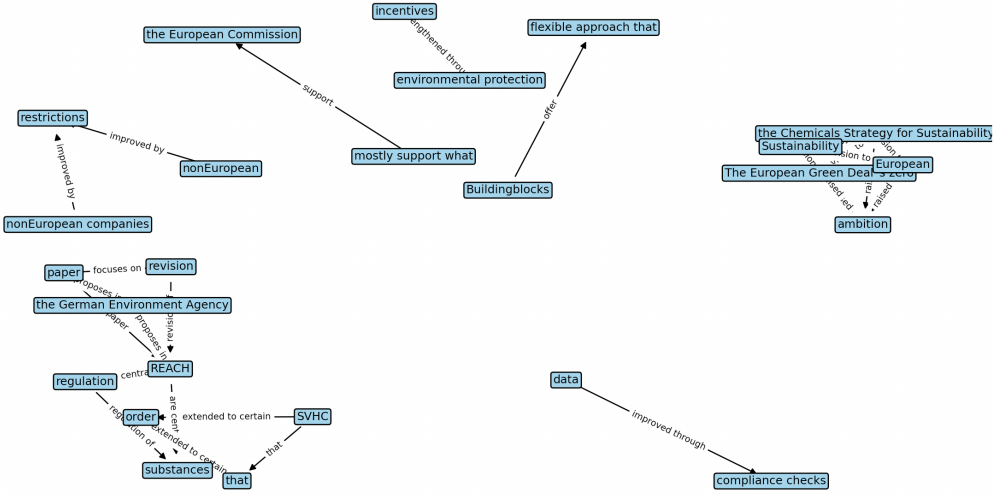
Knowledge graph provides interpretability

- Directed heterogeneous knowledge graph
- Makes text documents interpretable and understandable
- Several applications: paragraphs, single or multiple documents, filter entities etc.
- Not feasible by hand
- Create network from dataframe (networkx)

N-gram approach will improve with larger amount of data

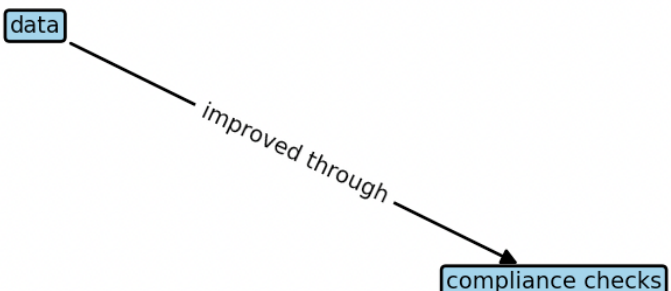


Root word approach finds meaningful relationships



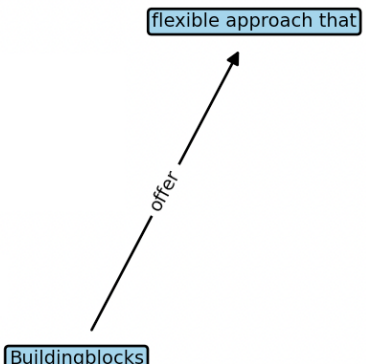
Example

"The availability and accessibility of data should be improved through compliance checks of all registration dossiers and additional information requirements for certain hazard classes."



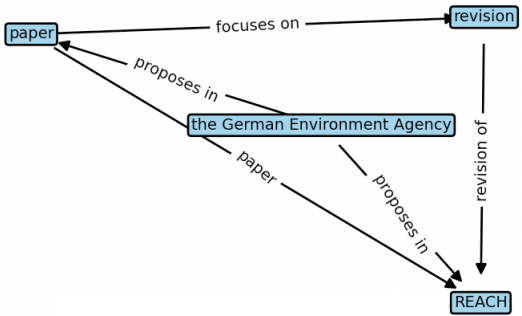
Example

"Building-blocks offer a flexible approach that can be adapted easily to different policy options."



Example

- "This paper focuses on the revision of the REACH authorisation and restriction system from an environmental perspective. [...] In conclusion, the German Environment Agency proposes in this paper a set of ambitious revisions to the REACH authorisation and restriction system."



Summary

- Data: Scientific papers from the German Environment Agency
- End-to-end pipeline: From PDF to Knowledge Graph
  - Combined named entity recognition with built rule-based approach
  - Adapted SelfORE model and build root-word approach for comparison
  - Generate both knowledge graphs for evaluation

- Results: Analysis of Examples

Bibliography

Xuming Hu et al. “SelfORE: Self-supervised Relational Feature Learning for Open Relation Extraction.” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3673–3682. DOI: 10.18653/v1/2020.emnlp-main.299.

Additional: German entity extraction

Neue exotische Stechmückenarten wie die Asiatische Tigermücke Aedes albopictus oder der **Japanische** Buschmoskito Aedes japonicus können als Vektoren für unterschiedliche **Viren** erheblich zur Ausbreitung neuer bisher in **Deutschland** nicht heimischer Infektionskrankheiten beitragen.
Seit 2007 wurden wiederholt einzelne Exemplare der Asiatischen Tigermücke und des **Japanischen** Buschmoskito in **Südwestdeutschland** nachgewiesen.
Bis 2013 hatte sich Ae. **japonicus** bereits flächendeckend in **BadenWürttemberg** sowie in großen Teilen von **NordrheinWestfalen** und **Niedersachsen** etabliert, während Ae. Aedes albopictus bis zu diesem Zeitpunkt nur sporadisch in **Deutschland** nachgewiesen wurde.
Als Haupteinfallspforte für Ae. **albopictuswaren** in einem früheren Forschungs und Entwicklungsvorhaben bundesdeutsche Autobahnen identifiziert worden, die einen starken Güter und Personenverkehr zu **südeuropäischen Ländern** aufwiesen.
Zur Entwicklung gezielter Präventionsmaßnahmen zum Schutz der Gesundheit von Mensch und Tier war es wichtig, den weiteren Eintrag sowie die mögliche Verbreitung und das **Etablierungsrisiko der Asiatischen Tigermücke** auch unter klimatischen Gesichtspunkten zu erfassen.

Additional: German root word Graphs