UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Klexikon: A German Dataset for Joint Summarization and Simplification

Dennis Aumiller and Michael Gertz
Heidelberg University
Institute of Computer Science
Database Systems Research Group
aumiller@informatik.uni-heidelberg.de

Outline

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

1. Reading a Simplified Article

2. Constructing the Klexikon Dataset

3. Exploratory Analysis

4. Future Work

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Reading a Simplified Article

• Reading a Simplified Article

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

- Expectation

Adler (Biologie)

Der Begriff **Adler** (von mittelhochdeutsch *adelar[e]* für ‚Edel-Aar', wobei *Aar* Adler und Adlerartige (unedle Adler) bezeichnete) ist im weiteren Sinn eine nicht genau definierte Sammelbezeichnung für große und kräftige Arten von Greifvögeln in der Familie der Habichtartigen, wie beispielsweise für die Gattung der Seeadler, den Schlangenadler oder den Philippinenadler. Im engeren Sinn steht die Bezeichnung *Adler* für die Gattung Echte Adler (Aquila) mit den ihr zugehörigen Arten. Deren größte lebende Vertreter sind der Steinadler und der Keilschwanzadler. Außerhalb der Habichtartigen wird im Deutschen auch der in seine eigene Familie gestellte Fischadler unter die Adler gezählt.


Kaiseradler

Mit einer Spannweite bis zu 3 Meter gilt der um 1400 n. Chr. ausgestorbene Haastadler (Harpagornis moorei) als größter bekannter Adler. Sein Lebensraum befand sich auf Neuseeland.

Jagd und Fortpflanzung [Bearbeiten | Quelltext bearbeiten]

Die außerordentliche Sehschärfe ihrer Augen gestattet es allen Adlern, ihre Beute auch aus großer Höhe zu erspähen und zu erjagen. Ihre Hauptbeute sind bodenlebende Wirbeltiere, in erster Linie Säuger. Nur selten ernähren sie sich von Aas. Adler sind sehr gut daran angepasst, zum Fliegen Aufwinde zu nutzen. Der Adler packt die Beute beim Niederstoßen aus dem Flug mit den Fängen. Dabei weisen drei seiner Zehen nach vorne, die Hinterkralle durchsticht den Beutekörper in einer Zangenbewegung. Durch den Griff wird die Beute erstickt.

In das aus Zweigen gebaute Nest (Adlerhorst) legt das Weibchen im Regelfall zwei bis drei Eier. Häufig wird jedoch nur eines der Jungen aufgezogen. Alle Adlerarten sind geschützt, ihr Bestand ist durch menschlichen Einfluss gefährdet.

Kunst und Symbolik [Bearbeiten | Quelltext bearbeiten]

Die Kunst des Alten Orients verwendet den Adler in vielfachen Abwandlungen als Symboltier. Bereits 3000 v. Chr. findet sich das Mischwesen Löwenadler in mesopotamischen Darstellungen, Doppeladler sind aus dem Babylonien des 23. Jahrhunderts v. Chr. bekannt. Die Antike kennt unzählige Abbildungen des Adlers auf Vasen, Gegenständen des täglichen Gebrauchs, Münzen und Schmuckstücken sowie in der Architektur auf Reliefs, Akroterien und Giebeln.

Seit dem 5. Jahrhundert nach Christus fertigten die Goten große Adlerfibeln und Schnallen mit Adlerköpfen. Aus der romanischen Periode des Mittelalters sind mit Adlern verzierte figürliche Darstellungen von Kapitellen bekannt. Ebenfalls hervorzuheben sind die Adlerpulte christlicher Altarräume. In der Malerei erscheint der Adler häufig als Symbol für den Evangelisten Johannes.


Adler in der Kunst – Symbol der Herrschaft, Kraft und Heldentum

In den Überlieferungen, dem Mythos und dem Volksglauben der frühen Hochkulturen (Ägypten, Mesopotamien) sowie der Völker des Altertums gilt der Adler allgemein als Symbol der Herrschaft und des Göttlichen.[1] Verstärkt wurde die Symbolik durch die Darstellung von Mischwesen wie dem Greif oder den Harpyien, welche die Kraft, Stärke oder auch Gesinnung durch unterschiedliche Tierkörperteile hervorhoben. Im altindischen Rigveda überbringt der Garuda als schlangentötender Bote die Nachrichten der Götter, insbesondere von Vishnu. Der Sage nach wurde der sumerische König Gilgamesch von einem Adler gerettet. Für die antiken Griechen war der Adler das Symbol des obersten olympischen Gottes Zeus, im alten Rom war er das Zeichen der obersten römischen Gottheit Jupiter sowie der kaiserlichen Macht und in der Apotheose Zeichen der Göttlichkeit des Kaisers.

Bei vielen nordamerikanischen Indianerstämmen galten ihre Federn – meist als Kopfschmuck – als Zeichen der Tapferkeit, aus den Knochen fertigten sie Pfeifen.

Adler

Zum Text für Lese-Anfänger

Adler sind große Greifvögel. Es gibt mehrere Arten, wie zum Beispiel Steinadler, Seeadler oder Fischadler. Sie ernähren sich von kleinen und größeren Tieren. Sie greifen ihre Beute mit ihren starken Krallen im Flug, am Boden oder im Wasser.


Hier streiten sich Seeadler um Beute.

Adler bauen ihre Nester, die man Horste nennt, meist auf Felsen oder hohen Bäumen. Dort hinein legt das Weibchen ein bis vier Eier. Die Brutzeit beträgt je nach Art 30 bis 45 Tage. Die Küken sind anfangs weiß, ihr dunkles Federkleid wächst erst später. Nach ungefähr 10 bis 11 Wochen können die Jungen fliegen.

Die bekannteste Adlerart in Mitteleuropa ist der Steinadler. Seine Federn sind braun und seine ausgestreckten Flügel sind etwa zwei Meter breit. Er lebt vor allem in den Alpen und rund ums Mittelmeer, aber auch in Nordamerika und Asien. Der Steinadler ist sehr kräftig und kann Säugetiere jagen, die schwerer sind als er selbst. Meist fängt er Hasen und Murmeltiere, aber auch junge Rehe und Hirsche, manchmal auch Reptilien und Vögel.

• Reading a Simplified Article

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

- Reality

Adler (Biologie)

Der Begriff **Adler** (von mittelhochdeutsch *adelar[e]* für ‚Edel-Aar', wobei *Aar* Adler und Adlerartige (unedle Adler) bezeichnete) ist im weiteren Sinn eine nicht genau definierte Sammelbezeichnung für große und kräftige Arten von Greifvögeln in der Familie der Habichtartigen, wie beispielsweise für die Gattung der Seeadler, den Schlangenadler oder den Philippinenadler. Im engeren Sinn steht die Bezeichnung *Adler* für die Gattung Echte Adler (Aquila) mit den ihr zugehörigen Arten. Deren größte lebende Vertreter sind der Steinadler und der Keilschwanzadler. Außerhalb der Habichtartigen wird im Deutschen auch der in seine eigene Familie gestellte Fischadler unter die Adler gezählt.

Mit einer Spannweite bis zu 3 Meter gilt der um 1400 n. Chr. ausgestorbene Haastadler (Harpagornis moorei) als größter bekannter Adler. Sein Lebensraum befand sich auf Neuseeland.


Kaiseradler

Jagd und Fortpflanzung [Bearbeiten | Quelltext bearbeiten]

Die außerordentliche Sehschärfe ihrer Augen gestattet es allen Adlern, ihre Beute auch aus großer Höhe zu erspähen und zu erjagen. Ihre Hauptbeute sind bodenlebende Wirbeltiere, in erster Linie Säuger. Nur selten ernähren sie sich von Aas. Adler sind sehr gut daran angepasst, zum Fliegen Aufwinde zu nutzen. Der Adler packt die Beute beim Niederstoßen aus dem Flug mit den Fängen. Dabei weisen drei seiner Zehen nach vorne, die Hinterkralle durchsticht den Beutekörper in einer Zangenbewegung. Durch den Griff wird die Beute erstickt.

In das aus Zweigen gebaute Nest (Adlerhorst) legt das Weibchen im Regelfall zwei bis drei Eier. Häufig wird jedoch nur eines der Jungen aufgezogen. Alle Adlerarten sind geschützt, ihr Bestand ist durch menschlichen Einfluss gefährdet.

Kunst und Symbolik [Bearbeiten | Quelltext bearbeiten]

Die Kunst des Alten Orients verwendet den Adler in vielfachen Abwandlungen als Symboltier. Bereits 3000 v. Chr. findet sich das Mischwesen Löwenadler in mesopotamischen Darstellungen, Doppeladler sind aus dem Babylonien des 23. Jahrhunderts v. Chr. bekannt. Die Antike kennt unzählige Abbildungen des Adlers auf Vasen, Gegenständen des täglichen Gebrauchs, Münzen und Schmuckstücken sowie in der Architektur auf Reliefs, Akroterien und Giebeln.

Seit dem 5. Jahrhundert nach Christus fertigten die Goten große Adlerfibeln und Schnallen mit Adlerköpfen. Aus der romanischen Periode des Mittelalters sind mit Adlern verzierte figürliche Darstellungen von Kapitellen bekannt. Ebenfalls hervorzuheben sind die Adlerpulte christlicher Altarräume. In der Malerei erscheint der Adler häufig als Symbol für den Evangelisten Johannes.

In den Überlieferungen, dem Mythos und dem Volksglauben der frühen Hochkulturen (Ägypten, Mesopotamien) sowie der Völker des Altertums gilt der Adler allgemein als Symbol der Herrschaft und des Göttlichen.[1] Verstärkt wurde die Symbolik durch die Darstellung von Mischwesen wie dem Greif oder den Harpyien, welche die Kraft, Stärke oder auch Gesinnung durch unterschiedliche Tierkörperteile hervorhoben. Im altindischen Rigveda überbringt der Garuda als schlangentötender Bote die Nachrichten der Götter, insbesondere von Vishnu. Der Sage nach wurde der sumerische König Gilgamesch von einem Adler gerettet. Für die antiken Griechen war der Adler das Symbol des obersten olympischen Gottes Zeus, im alten Rom war er das Zeichen der obersten römischen Gottheit Jupiter sowie der kaiserlichen Macht und in der Apotheose Zeichen der Göttlichkeit des Kaisers.


Adler in der Kunst – Symbol der Herrschaft, Kraft und Heldentum

Bei vielen nordamerikanischen Indianerstämmen galten ihre Federn – meist als Kopfschmuck – als Zeichen der Tapferkeit, aus den Knochen fertigten sie Pfeifen.

Jagdweise [Bearbeiten | Quelltext bearbeiten]

Steinadler jagen meist in offenen oder halboffenen Landschaften im bodennahen Flug unter optimaler Ausnutzung jeglicher Deckung. Sie gleiten dabei dicht an Hängen entlang, über Kuppen und kleine Hügel und versuchen ihre Beute auf kurze Distanz zu überraschen. Häufig jagen sie auch von einem Ansitz aus. Die Beute greifen die Adler meist auf dem Boden oder im bodennahen Luftraum und töten sie mit den außerordentlich kräftigen Zehen und Krallen. Sehr große Beutetiere wie Kitze des Steinbocks oder junge Gämsen greifen sie am Kopf. Der Steinadler schlägt dabei seine Krallen durch die Schädeldecke in das Gehirn. In den wenigen beobachteten Fällen wurden diese großen Beutetiere innerhalb von Sekunden getötet.

Weniger häufig ist die Jagd im freien Luftraum; die Erbeutung von ziehenden Kormoranen ist jedoch zum Beispiel schon mehrfach beobachtet worden. In Anbetracht ihrer Größe bewegen sich Steinadler in der Luft außerordentlich wendig und schnell, so wurde mehrfach beobachtet, wie sich ein Steinadler im Flug auf den Rücken drehte und so zum Beispiel einen verfolgenden Kolkraben erbeutete. Steinadler können keine Kadaver im Flug tragen, deren Gewicht das eigene Körpergewicht übertrifft. Schwere Beutetiere zerteilen sie daher und deponieren portionsweise, oder sie fliegen den Kadaver über mehrere Tage an.

Nahrung [Bearbeiten | Quelltext bearbeiten]

Steinadler sind außerordentlich kräftig und sehr geschickt. Sie erbeuten regelmäßig Tiere, die erheblich schwerer sind als sie selbst. Das maximale Beutegewicht liegt bei etwa 15 Kilogramm. Es gibt nur einen dokumentierten Fall, in dem ein noch schwererer Sikahirsch erlegt wurde.[4] Angriffsversuche auf annähernd ausgewachsene Gämsen sind dokumentiert.[5][6] Die Verteidigungsstrategie der Gämsen besteht darin, hangabwärts zu springen und sich überschlagend zu rollen, was eine erhebliche Verletzungsgefahr für beide bedeutet.

Im Beutespektrum dominieren meist bodenbewohnende, kleine bis mittelgroße Säugetiere von Ziesel- bis Steinbockkitz-Größe, Vögel spielen meist nur eine kleinere Rolle. Meist bilden wenige Säugerarten den Hauptteil der Nahrung. Daneben erbeutet der Steinadler jedoch fast alle kleinen und mittelgroßen Säuger und Vögel, die im jeweiligen Gebiet vorkommen. Insbesondere im Süden des Verbreitungsgebietes frisst er auch regelmäßig Reptilien, dort lassen Steinadler ähnlich wie Bartgeier auch Landschildkröten auf Felsen fallen, um so deren harten Panzer zu zerbrechen. Insbesondere im Winter, regional aber auch im Sommer, spielt Aas eine wichtige Rolle in der Ernährung.

Im Schweizer Kanton Graubünden dominierten zur Brutzeit im Beutespektrum Alpenmurmeltiere mit 60,2 % aller Beutetiere, an zweiter Stelle folgten junge Gämsen mit 8,0 %. Danach folgten Schneehase, Alpenschneehuhn und Birkhuhn mit jeweils 5,2 %.[7] Im schweizerischen Alpenvorland bestand die Nestlings-Nahrung in 4 Revieren vor allem aus Feldhasen (36,2 % aller Beutetiere), danach folgten Hauskatzen (27,5 %), Rehkitze (14,1 %) und Haushühner (8,1 %).[8] Populationen im Zentralmassiv Frankreichs jagen hauptsächlich Wildkaninchen. In Schottland wurden je nach Region Hasenartige in 10,7 % bis 46,9 % aller im Sommer gefundenen Gewölle nachgewiesen. Weitere wichtige Beutetiere waren dort Schafe und Ziegen (in 0,6 bis 26,8 % aller Gewölle), Raufußhühner (5,4 bis 47,8 %) und Rothirsche (als Aas) (1,2 bis 22,3 %).[9]

Raumnutzung und Siedlungsdichte [Bearbeiten | Quelltext bearbeiten]

Trotz des großen Verbreitungsgebietes liegen bisher nur wenige Daten zur Größe des Aktionsraumes, also zu der von einem Brutpaar genutzten Fläche vor. Die festgestellten Werte schwanken je nach Lebensraum und Nahrungsangebot erheblich. Im Schweizer Kanton Graubünden betrug die Größe des Aktionsraumes nach Sichtbeobachtungen in 26 Revieren zwischen 29 und 88 km², im Mittel 53 km². In Idaho beflogen vier Steinadlerpaare ebenfalls nach Sichtbeobachtungen Flächen zwischen 11,6 und 49,0 km², im Mittel 32,8 km².[10] Methodisch bedingt stellen die hier dargestellten Werte vermutlich eher die Untergrenze der tatsächlichen Aktionsraumgrößen dar. Steinadler verteidigen ihren gesamten Aktionsraum ganzjährig vehement gegen Artgenossen, eine früher gelegentlich vorgenommene Trennung zwischen dem verteidigten Revier und der zur Nahrungssuche genutzten Fläche ist also nicht gerechtfertigt.

Limitations of Current Text Simplification Systems

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

- Simplification:
  - "Translate" text into a simplified version (sentence-by-sentence)
  - Length of input corresponds to length of output!

- Instead, we should consider simplification as **document-level summarization**!
  - Simultaneously compress article length and simplify texts

Limitations of Current Text Simplification Systems

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

- But what about training data?
  - Corpora mostly provide sentence-level alignments
- Document-level resources barely exist!
  - For English:
  - For German:

Resource	Aligned Articles	Avg. #Sentences Source	Avg. #Sentences Simple
(Kauchak, 2013)	59,775	64.52	8.46
(Xu et al., 2015)	1,130	49.59	51.27
(Battisti et al., 2020)	378	45.29	55.75
(Hewett and Stede, 2021)	978	10.12	43.54

Table 1: Existing document-aligned resources for simplification.

Contributions

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

- Frame document-level simplification as a joint summarization and simplification problem

- Contribute a large-scale (document-aligned) resource for German based on alignments between Wikipedia and a German children's encyclopedia

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Constructing the Klexikon Dataset

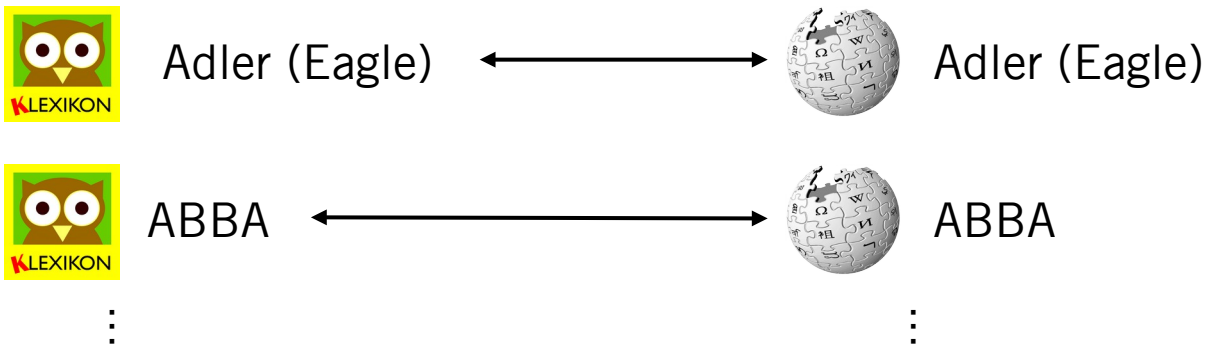
Klexikon

- Started in 2014 as a resource specifically for children (age 6-13)
- Each article is internally reviewed before release
- Almost 3,300 articles on diverse topics
  - Crawled during April 2021

Corpus Alignment

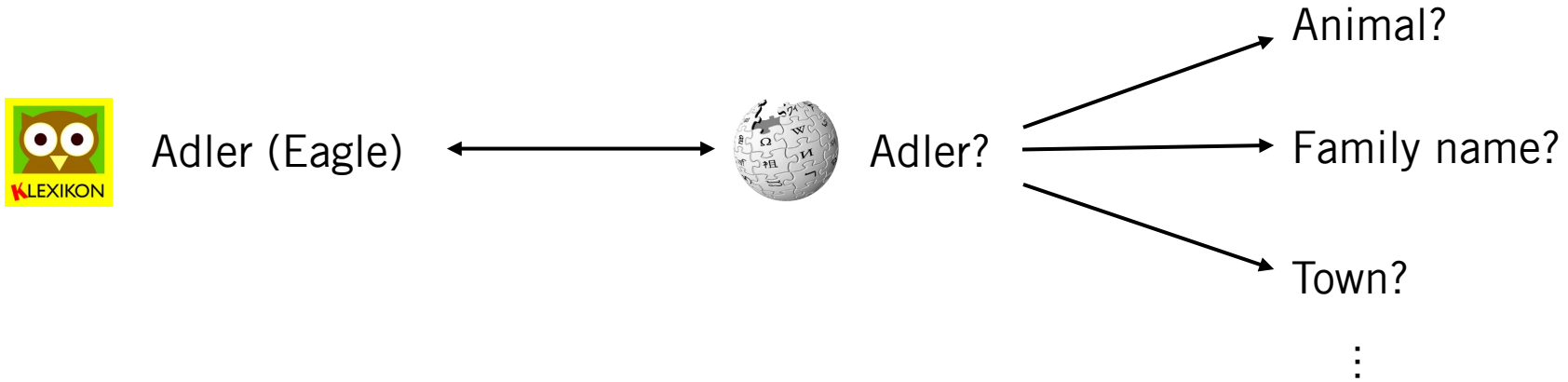
UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

- **Basic idea:** Klexikon articles are (summarized) simplifications of the Wikipedia article on the same topic
- Find Wikipedia article for each Klexikon entry (alignment)!
- **Hypothesis:** Aligned articles represent a suitable dataset for training a joint summarization/simplification objective



Corpus Alignment: Disambiguation

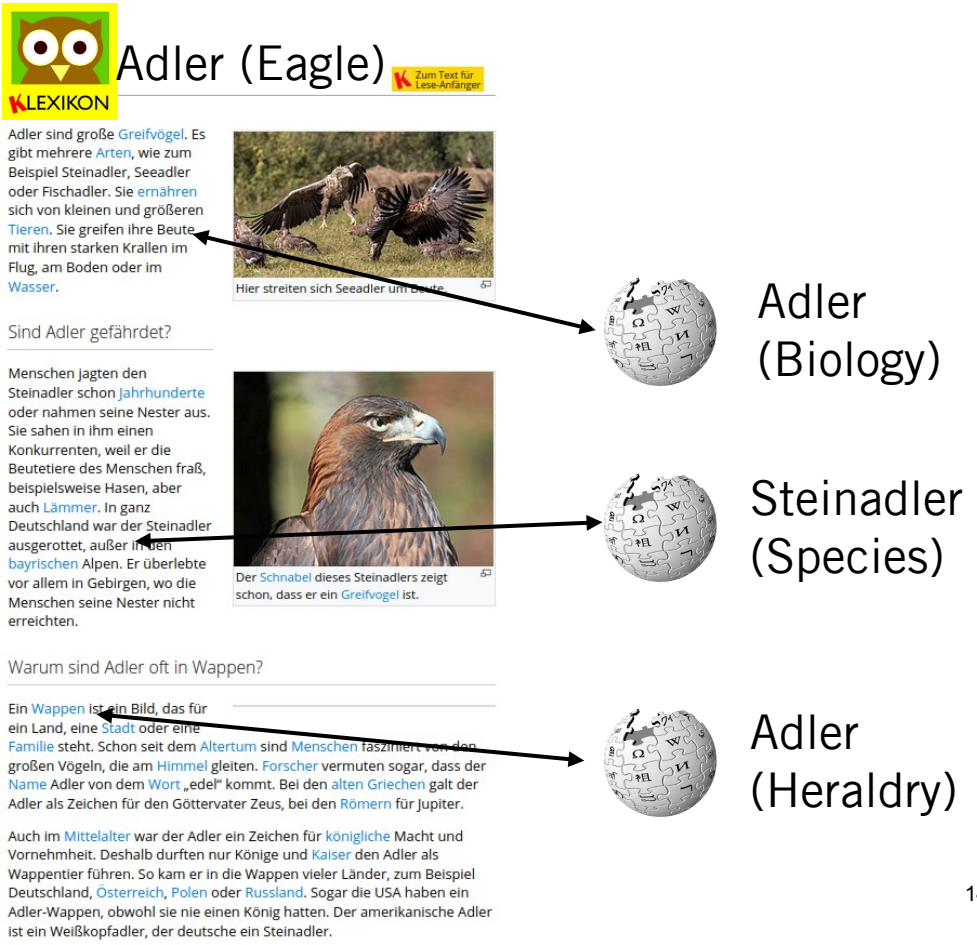
- Not always that easy: Only 90% match directly



- Requires manual resolution of conflicts

Corpus Alignment: Multi-Pages

- Content on Wikipedia can spread across pages
  - Manual review of disambiguated articles
  - Keep if more than 66% of the Klexikon article are corresponding to a *single* Wikipedia article



UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Exploratory Analysis

Resulting Corpus

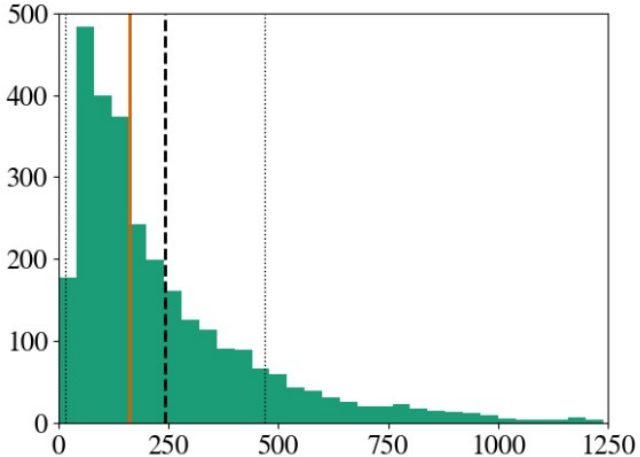
UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

- Around 2,900 articles aligned successfully!
- 3x more articles than largest previous corpus
- Significantly longer source documents

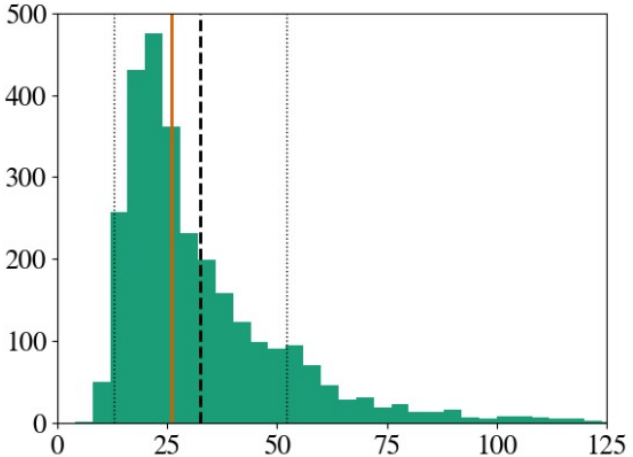
Resource	Aligned Articles	Avg. #Sentences Source	Avg. #Sentences Simple	Compr. Ratio
(Battisti et al., 2020)	378	45.29	55.75	0.81
(Hewett and Stede, 2021)	978	10.12	43.54	0.23
Klexikon (Ours)	**2,898**	242.09	32.51	**7.45**

Table 2: German resources in comparison.

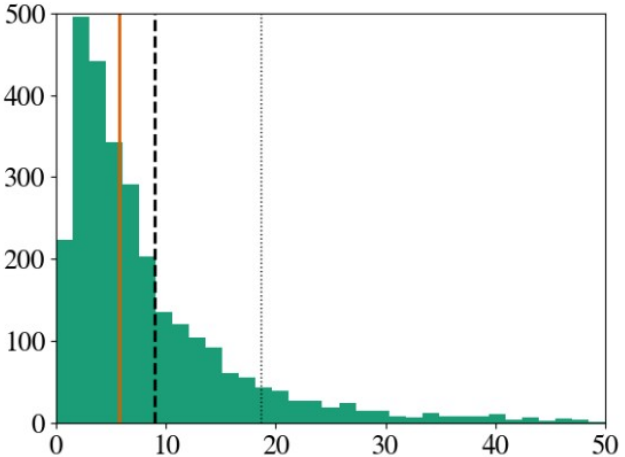
Length Distributions



Wikipedia article length (number of sentences)



Klexikon article length (number of sentences)



Compression ratio ($len_{wiki}/len_{klexikon}$)

Distribution of the dataset including median (red line), mean (dotted black) and standard deviation (thin black line) for both Wikipedia and Klexikon.

Suitability for Summarization

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

- How do we know the dataset is appropriate?
  - Test against (extractive) baselines!
- Compared methods:
  - Lead-3 (first three sentences)
  - Lead-k (opening paragraph in Wikipedia)
  - Full article
  - Luhn's Algorithm (Luhn, 1958)
  - LexRank (Erkan and Radev, 2004) (with sentence-transformers embeddings)
  - Extractive Oracle
- Evaluation with ROUGE

	R-1	R-2	R-L
Lead-3	16.95	3.77	9.81
Lead-k	24.87	5.10	12.01
Full article	16.81	4.23	6.95
Luhn	31.86	5.55	11.57
LexRank S-T	**33.90**	**6.11**	**12.86**
Oracle	41.85	10.68	16.00

Table 3: ROUGE F1 scores of baselines.

Suitability for Simplification

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

- Use modified Flesch score for German (Amstad, 1978)
- Average sentence length in tokens
- Average word length in characters
- Relative usage of the top 1000 lemmas of each corpus (Wikipedia/Klexikon) in relation to the total number of lemmas
  - Restrict to nouns, adjectives, verbs or adverbs
  - Less biased against longer texts

Suitability for Simplification

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

	Wikipedia	Klexikon
Mean Flesch score	40.1 ± 7.3	66.7 ± 6.0
Mean sentence length	22.7 ± 2.6	13.5 ± 1.5
Mean word length	8.7 ± 4.0	6.9 ± 3.0
Share of top 1000 lemmas	68.8%	82.3%

Table 4: Indicators of simplified texts across all metrics.

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Open Challenges and Future Work

Open Challenges

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

- No “sophisticated” system that performs both simplification and summarization exists yet
  - How to integrate simplification as part of summarizers?
  - Abstractive systems are limited by length
- How to align documents on a fine-grained context (sentence/paragraph-level) without manual annotation?
  - Existing solutions assume linear alignments and English

Future Work

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

- Provide additional sentence-level alignments
- Hybrid retrieval systems to cut down texts to suitable lengths for abstractive systems
- Experiment with regularization to incorporate simplification into neural networks

Resources

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Amstad, T. (1978). *Wie verständlich sind unsere Zeitungen?* Studenten-Schreib-Service.

Battisti, A., Pfütze, D., Säuberli, A., Kostrzewa, M., and Ebling, S. (2020). A corpus for automatic readability assessment and text simplification of German. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3302–3311, Marseille, France, May. European Language Resources Association.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Hewett, F. and Stede, M. (2021). Automatically evaluating the conceptual complexity of German texts. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 228–234, Düsseldorf, Germany, 6–9 September. KONVENS 2021 Organizers.

Kauchak, D. (2013). Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria, August. Association for Computational Linguistics.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Thank you for your attention!

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Check out the dataset on Huggingface:
https://huggingface.co/datasets/dennlinger/klexikon

The experimental code is also on Github:
https://github.com/dennlinger/klexikon

Any questions?

Or send a mail to aumiller@informatik.uni-heidelberg.de