DIS22: Komplexitätsbewertung englischer Texte - Datenannotiertung und Tool-Entwicklung im Projekt ARTS2 🎨2️⃣

Philipp Schaer, Björn Engelmann, Fabian Haak

Technology Arts Sciences
TH Köln

1. ARTS: Ursprung der Idee

Philipp Schaer, Björn Engelmann, Fabian Haak

Technology
Arts Sciences
TH Köln

Simpletext @ CLEF2023

Automatic Simplification of Scientific Texts for the General Public

*"The goal of this task is to provide a **simplified version of text passages**. Participants will be provided with the **popular science articles** and queries and matching abstracts of scientific papers."*

Philipp Schaer, Björn Engelmann, Fabian Haak

Technology
Arts Sciences
TH Köln

ChatGPT “2stepTurbo”

Indeed, modelling of an [infection network] between viral and cellular proteins will provide a conceptual and analytic framework to efficiently formulate new [biological hypothesis] at the [proteome] scale and to rationalize [drug discovery].

Philipp Schaer, Björn Engelmann, Fabian Haak

Von Simpletext zu ARTS

- Problem: Evaluation über “Readability Messung” und Ähnlichkeit zu “idealer Simplifizierung”
- Non-Experts können durchaus trotzdem komplexe Texte verstehen
- Idee: Evaluationsansatz entwickeln, der **verschiedene Anforderungen (Simplizitätsgrade) berücksichtigt**

BATS & ARTS

Evaluating Text Simplicity

Bewertung der Simplizität: Let’s Try it!

https://forms.gle/NQhu78RhmBhWv7h37

Philipp Schaer, Björn Engelmann, Fabian Haak

Technology
Arts Sciences
TH Köln

2. Was bedeutet simpel, was bedeutet komplex?

Philipp Schaer, Björn Engelmann, Fabian Haak

Technology
Arts Sciences
TH Köln

Text Simplification

- **(automatisierte) Simplifizierung von Texten**

Modelling of an infection network between viral and cellular proteins will provide a conceptual and analytic framework to efficiently formulate new biological hypothesis at the proteome scale and to rationalize drug discovery.

Building a model of infection pathways between viral and cellular components can help scientists generate new ideas for drugs at a large scale.

Was genau ist hier passiert? Warum ist der Text simpler?

Text Simplification

- **(automatisierte) Simplifizierung von Texten**

Modelling of an infection network between viral and cellular proteins will provide a conceptual and analytic framework to efficiently formulate new biological hypothesis at the proteome scale and to rationalize drug discovery.

Building a model of an infection network between viral and cellular components can help scientists generate new ideas for drugs at a large scale.

Text Simplification

Was macht schwierigen Text aus?

Philipp Schaer, Björn Engelmann, Fabian Haak

Technology
Arts Sciences
TH Köln

Text Simplification

- **Was macht simplen/schwierigen Text aus?**
  - **Lexikalische Aspekte (Fachwörter, Abkürzungen, Wortlänge, Satzlänge, …)**
  - **Syntaktische Aspekte (Grammatik; Tempora, Satzstruktur & Satzarten, …)**
  - **Nutzeraspekte**
  - **Domänenaspekte**

Text Simplification

Was könnte mit Domänen und Nutzeraspekten gemeint sein?

Philipp Schaer, Björn Engelmann, Fabian Haak

Technology
Arts Sciences
TH Köln

Text Simplification – User Aspects

A computer is an electronic device manipulating data through programmed instructions.

A computer is like a magic rock that can think really fast.

A computer is like a smart robot friend that helps you do all sorts of fun things!

Philipp Schaer, Björn Engelmann, Fabian Haak

Technology
Arts Sciences
TH Köln

Das Problem

Bisherige Ansätze betrachten Evaluation aus reiner Simplifizierungs-Perspektive:

Wie gut ist der Text simplifiziert worden?

Wir brauchen einen guten Ansatz, absolute Simplizität zu messen:

Wie simpel ist der Text generell?

Philipp Schaer, Björn Engelmann, Fabian Haak

Technology
Arts Sciences
TH Köln

Das Problem

Intuitiver Ansatz: Ein Modell mit Daten trainieren
ABER: Datensätze sind nicht geeignet, um Modelle zu entwickeln:

- Keine Label für Simplizität
- Simplifizierungen mancher Texte sind komplexer als andere komplexe Texte

Wir brauchen einen neuen Datensatz!
→ Wie lösen wir das Problem der Subjektivität und Spezifität?

3. Der ARTS Ansatz, euer Projekt

Philipp Schaer, Björn Engelmann, Fabian Haak

Bewertung der Simplizität: Let’s Try it *again*!

https://forms.gle/GQDQmwscpudRM8J69

Philipp Schaer, Björn Engelmann, Fabian Haak

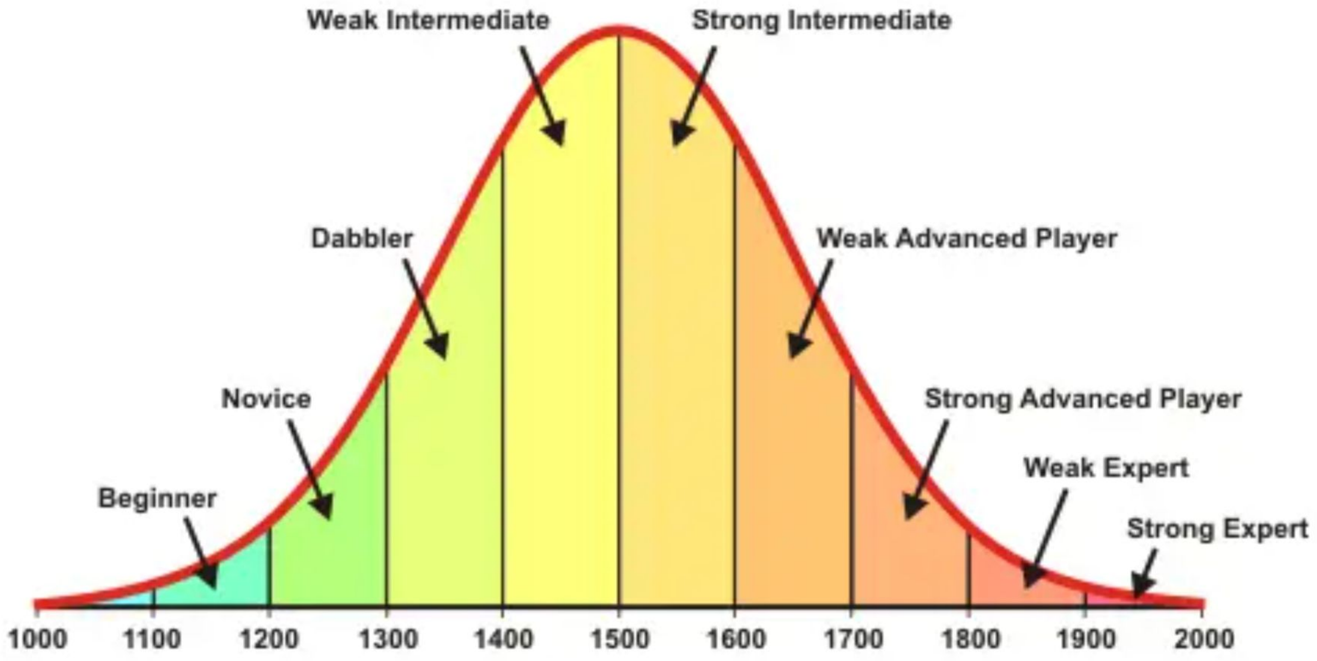
Technology
Arts Sciences
TH Köln

Die Lösung?

Direkte Quantifizierung ist schwer, entscheiden, welcher Text simpler ist (oft) leichter.

- über Vergleiche lassen sich Texte ranken
- je mehr Vergleiche von je mehr Personen, desto besser

Von den Vergleichen zu Scores



11/23/2015 9:31 AM	Shadow Fiend	4,876 (+25)	18:38 Ranked
11/23/2015 8:40 AM	Queen of Pain	4,851 (-25)	42:37 Ranked
11/23/2015 7:58 AM	Outworld Devourer	4,876 (+25)	34:44 Ranked
11/23/2015 7:14 AM	Queen of Pain	4,851 (+25)	33:38 Ranked
11/23/2015 6:09 AM	Templar Assassin	4,826 (+22)	53:49 Ranked
11/23/2015 4:59 AM	Slardar	4,804 (+25)	57:00 Ranked
11/23/2015 3:59 AM	Lina	4,779 (+25)	49:40 Ranked
11/23/2015 3:27 AM	Alchemist	4,754 (-24)	22:03 Ranked
11/23/2015 3:03 AM	Shadow Fiend	4,778 (+25)	14:23 Ranked
11/23/2015 1:46 AM	Alchemist	4,753 (+24)	44:14 Ranked
11/23/2015 12:49 AM	Earthshaker	4,729 (+25)	45:40 Ranked
11/23/2015 12:13 AM	Drow Ranger	4,704 (+26)	24:45 Ranked
11/22/2015 11:32 PM	Lina	4,678 (+25)	30:28 Ranked
11/22/2015 10:34 PM	Gyrocopter	4,653 (+22)	43:42 Ranked

Philipp Schaer, Björn Engelmann, Fabian Haak

Technology
Arts Sciences
TH Köln

Der ARTS Ansatz

1. **Über paarweise Vergleiche Texte gegeneinander antreten lassen**
2. **Über einen Elo-Algorithmus Matches bewerten**
3. **Aus Elo-Scores ein Ranking erstellen**
4. **Aus Ranking Simplicity Scores ableiten**

→ **Mit Datensatz Modelle trainieren!**

Das Projekt: Das Rating Interface

Philipp Schaer, Björn Engelmann, Fabian Haak

Technology
Arts Sciences
TH Köln

Das ARTS Projekt

Wir werden gemeinsam einen neuen ARTS Datensatz Labeln, der tiefer gehende Analysen erlaubt.

3 Task Forces:

1. **“Implis”: Neues Interface, neue Features**
2. **“Quantis”: Statistische Analyse der Daten**
3. **“Qualis”: Lassen sich intuitiv interessante Einsichten finden?**

https://github.com/irgroup/DIS22_ARTS_Project

Hausaufgabe

1. **ARTS Paper lesen**
2. **mindestens 3 Fragen überlegen:**
   a. Fragen zum Inhalt an uns
   b. Fragen zur Diskussion in der Gruppe

Nächste Woche werden wir darüber diskutieren!

→ Ihr bekommt das Paper per Mail

ARTS: Assessing Relative Text Simplicity 🎨

Anonymous ACL 2024 submission

Abstract

Automatic text simplification aims to reduce text complexity to make it better understandable for readers. Developing measures indicating the quality of text simplification approaches requires human-labeled data. Current measures quantifying simplicity of text suffer from the problem of being calibrated using *relative* data stemming from the question *"How much simpler has this text gotten compared to the original version?"* instead of the desirable *"How simple is the text?"*.

This work alleviates the problem by presenting datasets based on the question *"Which of two (unrelated) texts is simpler?"*. Using our **A**ssessing **R**elative **T**ext **S**implicity approach, we produce a general ranking. Through an Elo algorithm, we can

In most existing corpora with human labels on text simplicity, the simplicity is estimated as the agreement with a statement along the lines of *the simplification is easier to understand than the original text* (Alva-Manchego et al., 2020; Alva-Manchego et al., 2021; Maddela et al., 2023; Scialom et al., 2021; Sulem et al., 2018b). A rater thus is comparing a source text with the simplified version of the text and determining the *relative degree of simplification* that has been performed. As a distinction, raters do not assess the *absolute simplicity* of a text, but only how much it has improved. Development of measures that indicate the overall simplicity of a text requires datasets containing the combination of a text and a score quantifying its simplicity, reflecting the ease of reading and understanding.

Nächste Woche: Englisch Einstufung

Hintergrund: Wie hängen bestimmte Faktoren wie Syntax und Lexikalische Komplexität mit Englisch Niveau zusammen?

Durchgeführt auf Speexx (Selbstlernservice der TH Köln)

Wenn schon genutzt, Bescheid geben!

Datum	Dauer	Was steht an?
11.04.2024	90 min	Kick-Off, Vorstellung der Gruppenaufgaben
18.04.2024	90 min	Einteilung Aufgaben, **Englisch Einstufung**
25.04.2024	180 min*	**Vorstellung Zeitplan + Labeling Session**
02.05.2024	90 min	**Deadline Labeling Session**
09.05.2024		CHRISTI HIMMELFAHRT
16.05.2024	flexibel	Projektwoche (bei Bedarf: Implementationsgruppe)
23.05.2024	90 min	Flexible Meetingslots
30.05.2024		FRONLEICHNAM
06.06.2024	90 min	Vorstellung/Abnahme Implementation
13.06.2024	180 min*	**Präsentation Implementationsgruppe, Labeling Session 2**
20.06.2024	90 min	**Deadline Labeling Session 2**
27.06.2024	90 min	Flexible Meetingslots
04.07.2024	90 min	Flexible Meetingslots
11.07.2024	90 min	**Abschlusspräsentation**