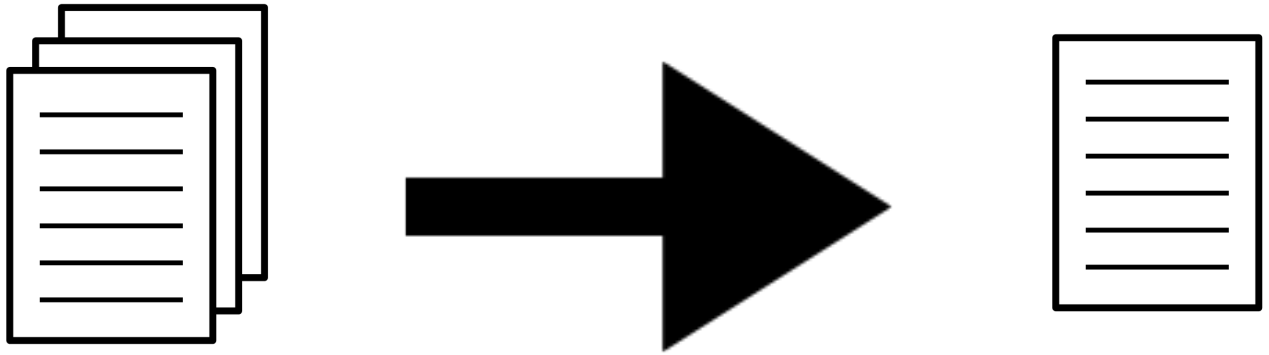
Automatic Detection of Linguistic Quality Violations

Jonathan Oberländer

Bachelor Thesis Defense
Universität des Saarlandes
21.08.2014

Automatic Summarization



Automatic Summarization

- **Single-Document:** One document
- **Multi-Document:** Multiple documents on the same topic

Automatic Summarization

- **Single-Document:** One document
- **Multi-Document:** Multiple documents on the same topic

- **Abstractive:** Internal semantic representation + generation
- **Extractive:** New summary from source sentences

Automatic Summarization

- **Single-Document:** One document
- **Multi-Document:** Multiple documents on the same topic

- **Abstractive:** Internal semantic representation + generation
- **Extractive:** New summary from source sentences

	Single-document	Multi-document
Abstractive		
Extractive		

Linguistic Quality Violations

Summarization systems should produce coherent and grammatical output.

Linguistic Quality Violations

Summarization systems don't produce coherent and grammatical output.
Why?

- It's hard.

Linguistic Quality Violations

Summarization systems don't produce coherent and grammatical output.
Why?

- It's hard.
- Evaluation: content, information density

Linguistic Quality Violations

Summarization systems don't produce coherent and grammatical output.
Why?

- It's hard.
- Evaluation: content, information density

$\Rightarrow$ LQVCorpus (Friedrich et al., 2014)

detecting ungrammaticality

Many subtypes of ungrammaticality

detecting ungrammaticality:subtypes

foo bar

detecting ungrammaticality:unknowntokens

foo bar

detecting ungrammaticality:unknowntokens

foo bar

detecting ungrammaticality:UT_evaluation

foo bar

detecting ungrammaticality:decision trees

foo bar

detecting ungrammaticality:evaluation

foo bar

detecting datelines

foo bar

detecting datelines:method

foo bar

detecting datelines:eval

foo bar

detecting redundancies

foo bar

detecting redundancies:method

foo bar

detecting redundancies:eval

foo bar

References

Friedrich, A., Valeeva, M., and Palmer, A. (2014). Lqvsumm: A corpus of linguistic quality violations in multi-document summarization.