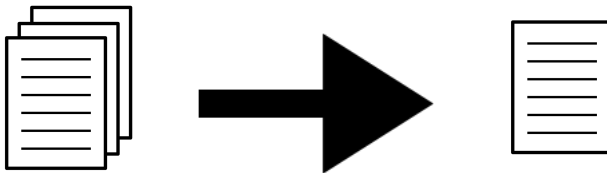


Automatic Detection of Linguistic Quality Violations

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		

Summarization systems should produce coherent and grammatical output.

Summarization systems **don't produce coherent and grammatical output.**

Why?

- ▶ It's hard.

Summarization systems **don't produce coherent and grammatical output.**

Why?

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

Summarization systems **don't produce coherent and grammatical output.**

Why?

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

⇒ LQVCorpus (Friedrich et al., 2014)

Annotated results of TAC 2011 Guided Summarization task
(Owczarzak and Dang, 2011)

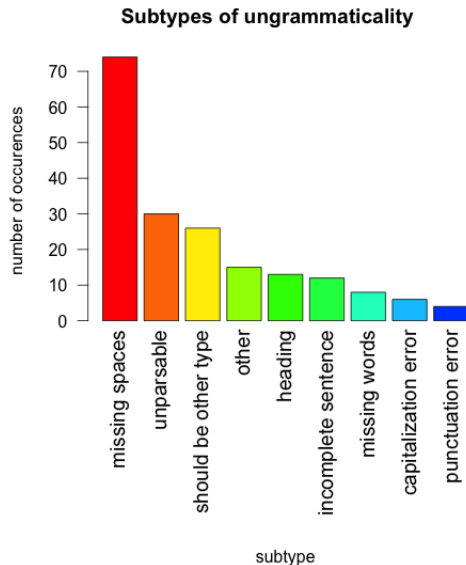
Annotated results of TAC 2011 Guided Summarization task
(Owczarzak and Dang, 2011)

- ▶ Entity level:
 - ▶ FM-EXPL, SM+EXPL
 - ▶ DNP-REF, INP+REF
 - ▶ PRN-ANT, PRN+MISLA
 - ▶ ACR-EXPL

Annotated results of TAC 2011 Guided Summarization task
(Owczarzak and Dang, 2011)

- ▶ Entity level:
 - ▶ FM-EXPL, SM+EXPL
 - ▶ DNP-REF, INP+REF
 - ▶ PRN-ANT, PRN+MISLA
 - ▶ ACR-EXPL
- ▶ Clause level:
 - ▶ **incomplete sentence (INCOMPLSN)**
 - ▶ **inclusion of datelines (INCLDATE)**
 - ▶ **other ungrammatical form (OTHRUNGR)**
 - ▶ no semantic relatedness (NOSEMREL)
 - ▶ **redundant information (REDUNINF)**
 - ▶ no discourse relation (NODISREL)

Ungrammaticality (OTHRUNGR+INCOMPLSN)



Detecting missing spaces

*“A strong earthquake measuring 7.8 magnitude struck **Wenchuancounty** of Sichuan Province on Monday, leaving at least **12,000people** died and thousands more injured.”*

*“Virginia Tech reported a campus shooting Monday and told **studentsto** stay inside their residences and away from windows.”*

*“A gunman opened fire on classrooms at Virginia Tech University **onMonday** morning, killing at least 30 people before turning his **gunon** himself in the bloodiest school shooting in US history.”*

Idea:

Sentence contains violation iff any word \notin known tokens

Idea:

Sentence contains violation iff any word \notin known tokens

known tokens?

- ▶ Source documents available? \rightarrow all tokens in source documents = **UnknownTokens**_{source}
- ▶ Otherwise \rightarrow **UnknownTokens**_{general}

- ▶ Tokens from (parts of) Gigaword = **UnknownTokens_{gw}**
- ▶ + Heuristics (Capitalized words) = **UnknownTokens_{gw+heur}**
- ▶ + NER (Finkel et al., 2005) = **UnknownTokens_{gw+heur+ner}**
- ▶ + Wikipedia = **UnknownTokens_{gw+heur+ner+wiki}**

foo bar

detecting ungrammaticality:decision trees

foo bar

foo bar

foo bar

foo bar

foo bar

foo bar

foo bar

foo bar

- Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.
- Friedrich, A., Valeeva, M., and Palmer, A. (2014). Lqvsumm: A corpus of linguistic quality violations in multi-document summarization.
- Owczarzak, K. and Dang, H. T. (2011). Overview of the tac 2011 summarization track: Guided task and aesop task. In *Proceedings of the Text Analysis Conference (TAC 2011)*, Gaithersburg, Maryland, USA, November.