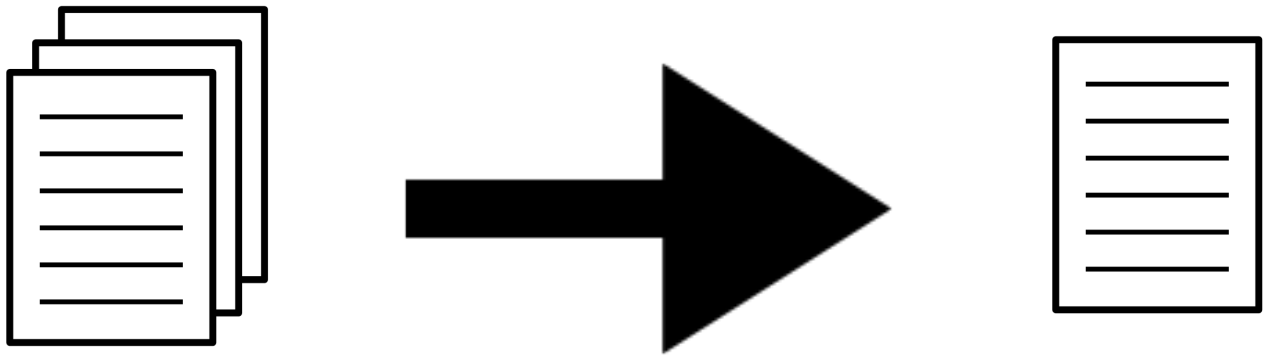
Automatic Detection of Linguistic Quality Violations

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)

LQVCorpus (Friedrich et al., 2014)

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

LQVCorpus (Friedrich et al., 2014)

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

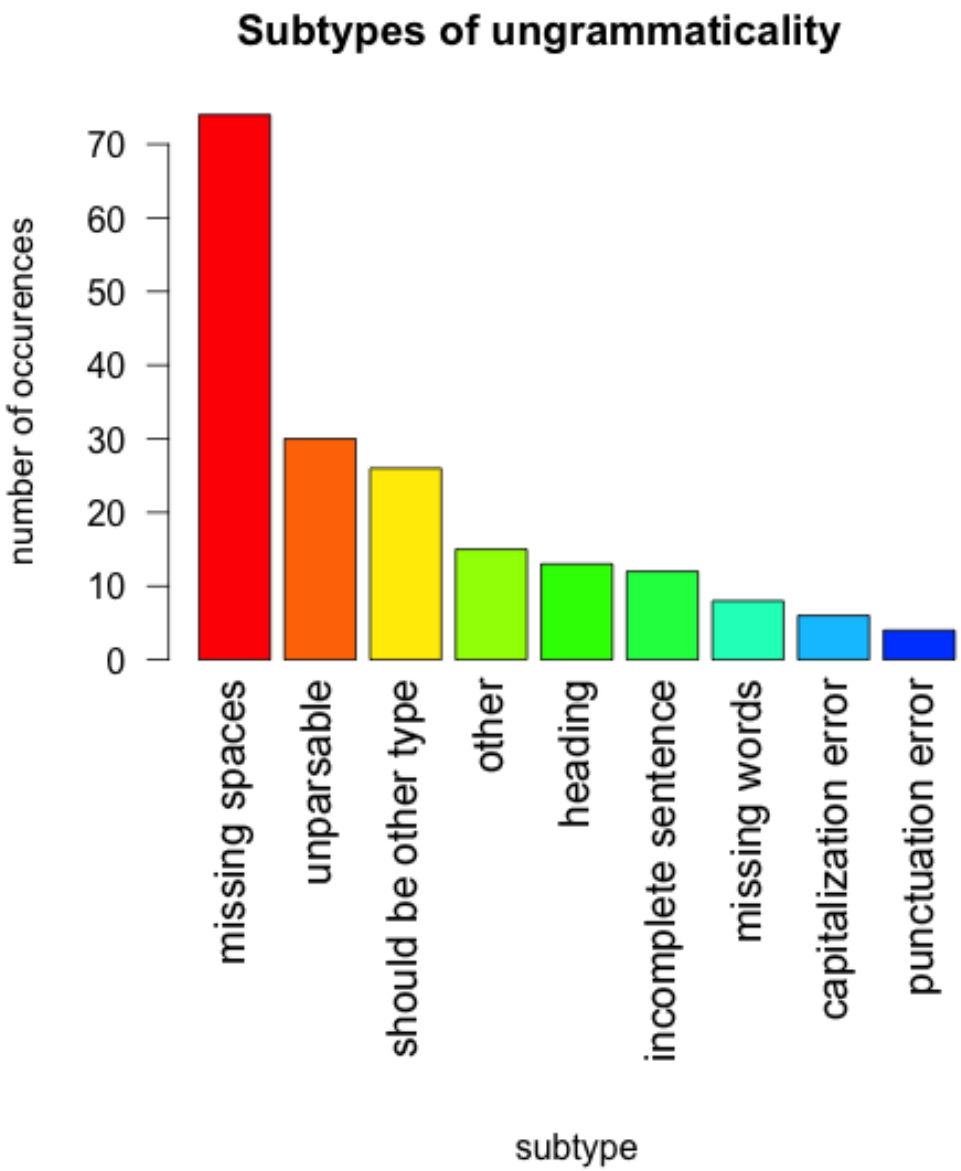
- Entity level:
  - FM-EXPL, SM+EXPL
  - DNP-REF, INP+REF
  - PRN-ANT, PRN+MISLA
  - ACR-EXPL

LQVCorpus (Friedrich et al., 2014)

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."*

UnknownTokens

Idea:
Sentence contains violation iff any word $\notin$ known tokens

UnknownTokens

Idea:
Sentence contains violation iff any word $\notin$ known tokens

known tokens?

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

$\mathbf{UnknownTokens}_{\mathbf{general}}$

- Tokens from (parts of) Gigaword = $\mathbf{UnknownTokens}_{\mathbf{gw}}$

$\textbf{UnknownTokens}_{\textbf{general}}$

- Tokens from (parts of) Gigaword = $\textbf{UnknownTokens}_{\textbf{gw}}$
- + Heuristics (Capitalized words) = $\textbf{UnknownTokens}_{\textbf{gw+heur}}$

$\textbf{UnknownTokens}_{\textbf{general}}$

- Tokens from (parts of) Gigaword = $\textbf{UnknownTokens}_{\textbf{gw}}$
- + Heuristics (Capitalized words) = $\textbf{UnknownTokens}_{\textbf{gw+heur}}$
- + NER (Finkel et al., 2005) = $\textbf{UnknownTokens}_{\textbf{gw+heur+ner}}$

$\textbf{UnknownTokens}_{\textbf{general}}$

- Tokens from (parts of) Gigaword = $\textbf{UnknownTokens}_{\textbf{gw}}$
- + Heuristics (Capitalized words) = $\textbf{UnknownTokens}_{\textbf{gw+heur}}$
- + NER (Finkel et al., 2005) = $\textbf{UnknownTokens}_{\textbf{gw+heur+ner}}$
- + Wikipedia = $\textbf{UnknownTokens}_{\textbf{gw+heur+ner+wiki}}$

UnknownTokens: Evaluation

	Missing spaces			No missing spaces		
	P	**R**	**F**	**P**	**R**	**F**
Baseline	0.0	0.0	0.0	94.8	**100**	97.3
UT_{gw}	15.0	**98.7**	26.0	**99.9**	69.1	81.7
$UT_{gw+heur}$	30.5	97.3	46.5	99.8	87.8	93.4
$UT_{gw+heur+ner}$	35.5	97.3	52.0	99.8	90.3	94.8
$UT_{gw+heur+ner+wiki}$	70.3	96.0	81.2	99.8	97.8	98.8
UT_{source}	**95.9**	94.6	**95.2**	99.7	99.7	**99.7**

RandomForest

RandomForest (Breiman, 2001) to train decision trees

Features:

- classification from **UnknownTokens**

RandomForest

RandomForest (Breiman, 2001) to train decision trees

Features:

- classification from **UnknownTokens**
- perplexity scores from language model trained on Gigaword

RandomForest

RandomForest (Breiman, 2001) to train decision trees

Features:

- classification from **UnknownTokens**
- perplexity scores from language model trained on Gigaword
- number of words

RandomForest: Evaluation

	Precision	**Recall**	**F-Score**
Ungrammatical	59.6	29.3	39.3
Not ungrammatical	90.1	97.0	93.4
Weighted Average	86.1	88.1	86.3

Datelines

BLACKSBURG, Virginia 2007-04-16 18:34: 44 UTC *A gunman opened fire in a dorm and classroom at Virginia Tech on Monday, killing at least 30 people in the deadliest shooting rampage in U.S. history.*

BERLIN, May 13(Xinhua) *The German government announced on Tuesday that it is to provide 500, 000 euros(around 770, 000 U.S. dollars) in aid for earthquake victims in Sichuan Province of China.*

00 a.m.*People are panicking.*

Detecting Datelines

Regular expression:

```
UTC|
^\d{4}-\d{2}-\d{2}|
^[A-Z]{3,}|
^(Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)
```

Precision	**Recall**	**F-Score**
90.3%	91.1 %	90.7%

Datelines Evaluation

foo bar

detecting redundancies

foo bar

detecting redundancies:method

foo bar

detecting redundancies:eval

foo bar

References

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

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*.