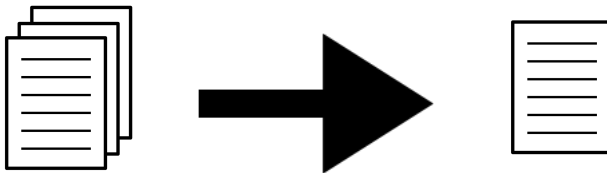


Automatic Detection of Linguistic Quality Violations

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:
 - ▶ definite noun phrase without reference to previous mention
 - ▶ pronoun with missing antecedent
 - ▶ acronym without explanations
 - ▶ ...

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

- ▶ Entity level:
 - ▶ definite noun phrase without reference to previous mention
 - ▶ pronoun with missing antecedent
 - ▶ acronym without explanations
 - ▶ ...
- ▶ Clause level:
 - ▶ **incomplete sentence (INCOMPLSN)**
 - ▶ **inclusion of datelines (INCLDATE)**
 - ▶ **other ungrammatical form (OTHRUNGR)**
 - ▶ no semantic relatedness (NOSEMREL)
 - ▶ **redundant information (REDUNINF)**
 - ▶ no discourse relation (NODISREL)

- ▶ small
- ▶ partially inconsistent

- ▶ small
- ▶ partially inconsistent
- ▶ corpus preprocessing with CoreNLP (Manning et al., 2014)
- ▶ unit of annotation (clauses vs. sentences)

- ▶ small
- ▶ partially inconsistent
- ▶ corpus preprocessing with CoreNLP (Manning et al., 2014)
- ▶ unit of annotation (clauses vs. sentences)
- ▶ OTHRUNGR has different violation subtypes

Development and Test Sets

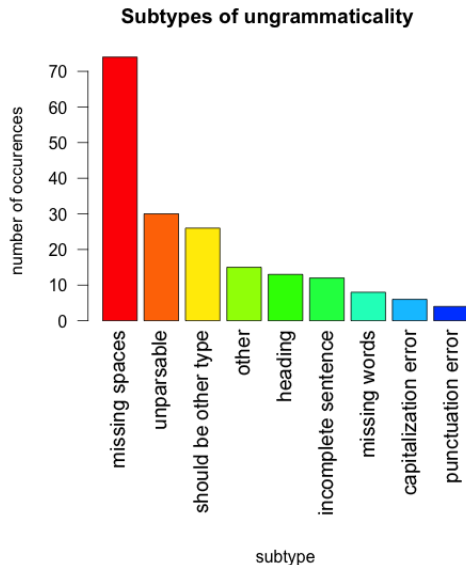
2 development sets:

- ▶ *dev-1*: 20% (D1101-D1108)
- ▶ *dev-2*: 20% (D1109-D1116)

1 test set:

- ▶ *test*: 60% (D1117-D1144)

Ungrammaticality (OTHRUNGR+INCOMPLSN) on *dev-2*



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.”*

Unknown Tokens

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 = **UnknownTokens**_{source}
- ▶ Otherwise \rightarrow **UnknownTokens**_{general}

- ▶ Tokens from (parts of) Gigaword = **UnknownTokens_{gw}**

- ▶ Tokens from (parts of) Gigaword = **UnknownTokens_{gw}**
- ▶ + Heuristics (Capitalized words) = **UnknownTokens_{gw+heur}**

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

- ▶ 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}**

UnknownTokens: Evaluation on *dev-2*

	Missing spaces		
	P	R	F
UT_{source}	95.9	94.6	95.2

UnknownTokens: Evaluation on *dev-2*

	Missing spaces		
	P	R	F
UT_{source}	95.9	94.6	95.2
UT_{gw}	15.0	98.7	26.0

UnknownTokens: Evaluation on *dev-2*

	Missing spaces		
	P	R	F
UT_{source}	95.9	94.6	95.2
UT_{gw}	15.0	98.7	26.0
UT_{gw+heur}	30.5	97.3	46.5

UnknownTokens: Evaluation on *dev-2*

	Missing spaces		
	P	R	F
UT_{source}	95.9	94.6	95.2
UT_{gw}	15.0	98.7	26.0
UT_{gw+heur}	30.5	97.3	46.5
UT_{gw+heur+ner}	35.5	97.3	52.0

UnknownTokens: Evaluation on *dev-2*

	Missing spaces		
	P	R	F
UT_{source}	95.9	94.6	95.2
UT_{gw}	15.0	98.7	26.0
$UT_{gw+heur}$	30.5	97.3	46.5
$UT_{gw+heur+ner}$	35.5	97.3	52.0
$UT_{gw+heur+ner+wiki}$	70.3	96.0	81.2

RandomForest (Breiman, 2001) to train decision trees

Features:

- ▶ classification from **UnknownTokens**

RandomForest (Breiman, 2001) to train decision trees

Features:

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

RandomForest (Breiman, 2001) to train decision trees

Features:

- ▶ classification from **UnknownTokens**
- ▶ perplexity scores from language model trained on Gigaword
- ▶ number of words
- ▶ 3 features from ACE parser output

RandomForest: Evaluation on *test*

	Precision	Recall	F-Score
Ungrammatical	72.8	49.1	58.6

Ablation Study:

Feature	Decrease in Accuracy
UnknownTokens Output	0.3369
Language Model Perplexity	1.2091
Number of Words	0.7334
ACE RAM	1.0901
ACE Readings	1.3478
ACE Status	0.1982

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)

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)
```

Evaluation on *test*:

Precision	Recall	F-Score
86.0%	89.7%	87.8%

*According to a survey by the State **Food and Drug Administration**, 65 percent of the respondents worried about the food **safety** situation in China.*

***Food and drug safety** has become a major concern of Chinese people.*

***Cyclone Sidr**, described as the worst storm in years to hit low-lying and disaster-prone **Bangladesh**, crashed into the southwestern coast **Thursday night** before sweeping north over the capital **Dhaka**.*

*The **cyclone** hit the southwestern coast of **Bangladesh** on **Thursday** before sweeping north to the capital **Dhaka**.*

Mary saw the 5 “elephants”. She saw the horses.

$\{Mary, saw, the, 5, elephants\}, \{She, saw, the, horses\}$

- ▶ Remove non-alphanumeric characters and split into set of words

$$|\{saw, the\}| = 2$$

- ▶ Remove non-alphanumeric characters and split into set of words
- ▶ Cardinality of intersection between sets

$$score = \frac{2}{|\{She, saw, the, horses\}|} = 0.5$$

- ▶ Remove non-alphanumeric characters and split into set of words
- ▶ Cardinality of intersection between sets
- ▶ Normalize by sentence length

$0.5 > \textit{threshold} ?$

- ▶ Remove non-alphanumeric characters and split into set of words
- ▶ Cardinality of intersection between sets
- ▶ Normalize by sentence length
- ▶ Classify with threshold

$0.5 > \text{threshold} ?$

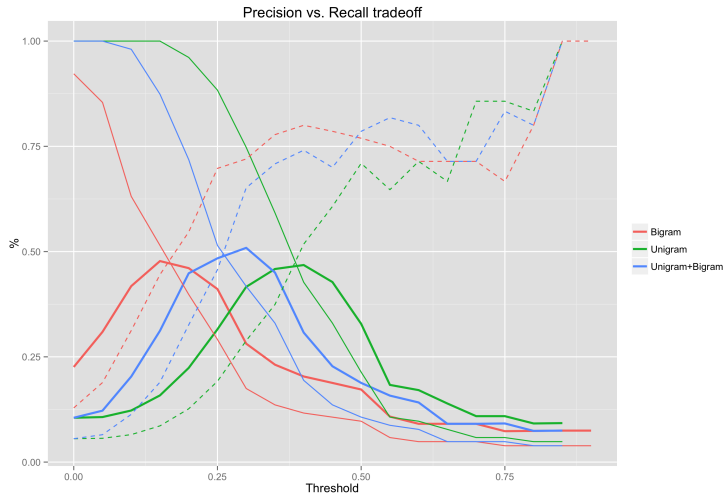
- ▶ Remove non-alphanumeric characters and split into set of words
- ▶ Cardinality of intersection between sets
- ▶ Normalize by sentence length
- ▶ Classify with threshold

Variations: **Bigrams**, **Combined**

Threshold?

Finding a threshold

dev-1+2



Evaluation of **Unigrams**, ... on *test*

	Unigrams	Bigrams	Combined
Threshold	0.5	0.4	0.4

	Precision	Recall	F-Score
Baseline	4.5%	100%	8.7%
Levenshtein	15.8%	17.3%	3.1%
Unigrams	58.0%	28.2%	37.0%
Bigrams	55.6%	14.5%	22.9%
Combined	56.8%	24.3%	34.0%

Methods for:

- ▶ detecting ungrammaticality
- ▶ detecting datelines
- ▶ detecting redundancies

Methods for:

- ▶ detecting ungrammaticality
- ▶ detecting datelines
- ▶ detecting redundancies

Adapted annotation scheme, better for automatic processing

Methods for:

- ▶ detecting ungrammaticality
- ▶ detecting datelines
- ▶ detecting redundancies

Adapted annotation scheme, better for automatic processing

Tool will be made available to annotate with our methods

Other violations

- ▶ pronouns: coreference resolution
- ▶ acronyms: finding full form near first unexpanded form
- ▶ mentions & noun phrases: NER + ?
- ▶ no semantic relatedness: semantic parsing? Wordnet distance?
- ▶ no discourse relation: discourse parsing, does connective match relation?

Ungrammaticality

- ▶ detection methods for other subtypes

Ungrammaticality

- ▶ detection methods for other subtypes

Redundancy

- ▶ include contextual information
- ▶ include source document information
- ▶ semantic approaches

Ungrammaticality

- ▶ detection methods for other subtypes

Redundancy

- ▶ include contextual information
- ▶ include source document information
- ▶ semantic approaches

Corpus

- ▶ annotate a corpus with subtypes, sentence based
- ▶ evaluate methods on other data sets/corpora/domains

- 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.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.
- 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*.

Bonus Slide: Full **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

Bonus Slide: **RandomForest**: Evaluation of all classes

	Precision	Recall	F-Score
Ungrammatical	72.8	49.1	58.6

Bonus Slide: **RandomForest**: Evaluation of all classes

	Precision	Recall	F-Score
Ungrammatical	72.8	49.1	58.6
Not ungrammatical	86.6	94.7	90.5
Weighted Average	83.5	84.5	83.4