LQVSumm: A Corpus of Linguistic Quality Violations in Multi-Document Summarization

Annemarie Friedrich, Marina Valeeva and Alexis Palmer

Department of Computational Linguistics
Saarland University, Saarbrücken, Germany
(afried|marinav|apalmer)@coli.uni-saarland.de

Abstract

We present LQVSumm, a corpus of about 2000 automatically created extractive multi-document summaries from the TAC 2011 shared task on Guided Summarization, which we annotated with several types of linguistic quality violations. Examples for such violations include pronouns that lack antecedents or ungrammatical clauses. We give details on the annotation scheme and show that inter-annotator agreement is good given the open-ended nature of the task. The annotated summaries have previously been scored for Readability on a numeric scale by human annotators in the context of the TAC challenge; we show that the number of instances of violations of linguistic quality of a summary correlates with these intuitively assigned numeric scores. On a system-level, the average number of violations marked in a system's summaries achieves higher correlation with the Readability scores than current supervised state-of-the-art methods for assigning a single readability score to a summary. It is our hope that our corpus facilitates the development of methods that not only judge the linguistic quality of automatically generated summaries as a whole, but which also allow for detecting, labeling, and fixing particular violations in a text.

Keywords: multi-document summarization, linguistic quality evaluation, discourse coherence

1. Introduction

While automatic summarization systems are getting better at covering the most important content of the documents or document collections they summarize, the linguistic quality of automatically generated summaries leaves much room for improvement (Nenkova and McKeown, 2011). Neither do we have entirely adequate methods for automatically evaluating linguistic quality (LQ) of summaries.

The work in this paper specifically addresses the task of LQ evaluation for *extractive multi-document summarization*, where summaries are generated by extracting sentences or clauses from a collection of documents. State-of-the-art extractive summarization systems often apply sentence compression, e.g. by removing phrases, such that the resulting summaries are usually, but not always, fully grammatical. As this work shows, other frequent shortcomings of the LQ of the generated summaries, however, are related to discourse phenomena such as coreference problems.

Previous works on LQ evaluation (Conroy et al., 2011; Giannakopoulos and Karkaletsis, 2011; de Oliveira, 2011; Pitler et al., 2010; Lin et al., 2012) all use automatically obtainable lexical, syntactic and/or semantic features to create numeric scores for summaries in supervised settings and achieve promising results (more details given in section 5). One disadvantage of these approaches is that it remains unclear precisely which aspects of LQ contribute most to the readability of summaries; we hypothesize that some violations have more impact than others. A second disadvantage is that the methods output a score for LQ without detecting and labeling particular instances of violations.

Figure 1 shows an example summary produced by an extractive summarization system, chosen to illustrate several LQ violations. First, it is inappropriate to use a definite article when an entity unknown to the reader is mentioned for the first time (*the girls, the Amish schoolhouse*). Other types of first reference to previously-unknown entities also need explanation: e.g. who is *Miller*? The detailed explanation *a local truck driver Charles Roberts* is also misplaced, as Roberts has already been introduced to the discourse. Finally, the passages marked in bold indicate redundant information within the summary.

> *Charles Carl Roberts IV may have planned to molest the girls at the Amish school, but police have no evidence that he actually did. Charles Carl Roberts IV entered the West Nickel Mines Amish School in Lancaster County and shot 10 girls, killing five. The suspect apparently called his wife from a cell phone shortly before the shooting began, saying* ***he was "acting out in revenge for something that happened 20 years ago****, Miller said. The gunman, a local truck driver Charles Roberts, was apparently* ***acting in "revenge" for an incident that happened to him 20 years ago.***

Figure 1: Automatically created summary from the TAC 2011 Guided Summarization task.[1]

With this paper, we introduce LQVSumm, a collection of annotations of specific LQ violations in automatically-produced extractive summaries.[2] We develop an annotation scheme for such violations (section 2) and annotate 1,985 summaries from the TAC 2011 Guided Summariza-

[1] http://www.nist.gov/tac/2011/Summarization

[2] The corpus is available in a stand-off format via the LREC META-SHARE and from http://www.coli.uni-saarland.de/~afried.

tion Task (Owczarzak and Dang, 2011) and 50 summaries generated by G-Flow (Christensen et al., 2013), which aims specifically at creating summaries optimized for coherence. For example, among other types of violations, we mark pronouns that lack antecedents, adjacent sentences that are not semantically related, and ungrammatical sentences.

In section 3, we give the details of an inter-annotator agreement study, which shows that inter-annotator agreement is substantial for annotations on the clause level, and acceptable for annotations on the level of entity mentions. We also present an overview of the corpus annotations, revealing the most frequent LQ errors made by state-of-the-art summarization systems: definite noun phrases without a previous reference to the same entity, first mentions of entities without a clear reference, incomplete and ungrammatical sentences, just to name the top of the list. We also investigate the statistical relationship between the LQ violations annotated in a summary and its manually (intuitively) assigned *Readability* score assigned by the judges of the TAC challenge (section 4), finding that almost all violation types in our annotation scheme have an influence on readability scores. Further, we find that using the number of (gold-standard) LQ violations to rank summarization systems outperforms a current state-of-the-art supervised method for this task (Lin et al., 2012).

These are promising results, which show that detecting LQ violations is a suitable method for evaluating the LQ of summaries and, further, that a system able to reliably detect LQ violations could outperform state-of-the-art automatic methods for LQ evaluation. The corpus and analysis presented here facilitate the development of such methods, which we aim to address in future work. It is our hope that LQVSumm will open up new possibilities for research in the area of linguistic quality evaluation and the development of summarization systems aimed at producing coherent summaries of high linguistic quality.

2. Annotation Scheme

To identify relevant violations of linguistic quality, we first manually inspected some of the extractive summaries provided by TAC (see section 3). We consider two classes of LQ violations. First, many LQ violations involve problems with reference or coreference; these are marked at the level of entity mentions (section 2.1). Other violations such as ungrammaticality or redundancy take larger scope; these are marked at the level of clauses (section 2.2). In this section, we present the details of our annotation scheme, developed to mark various categories of LQ violations.[3]

2.1. LQ violations on the level of entity mentions

We annotate *entity mentions*, i.e., common nouns, named entities and pronouns, that are involved in violations of coherence or readability. In the following, we list and explain the various types of violations annotated at this level; these are realized as features on the entity mentions.

[3]All examples below are taken from summaries of the TAC 2011 Guided Summarization Task, for which we provide annotations in LQVSumm.

first mention without explanation (FM-EXPL): First mentions of entities within a discourse are somehow special; a reader unfamiliar with the events reported in the text must be able to determine the referent for newly-mentioned entities. This feature is assigned to first mentions of an entity that lack a clear reference for the reader. In example (1), which is the first sentence of a summary, *'Roberts'* lacks sufficient explanation and is hence assigned this feature.

(1) *Roberts killed himself in the one-room remote Amish schoolhouse before police could get to him.*

Well known entities (*'President Obama'*) or entities that are introduced with a short description (*'Tony Taylor, 34, of Hampton, Va.'*) are not marked with this feature.

subsequent mention with explanation (SM+EXPL): This feature marks mentions of entities that have already been referenced in the text but still appear with an inappropriately explanatory introduction. In example (2), *'Tony Taylor'* is assigned this feature, and additionally a link between the overly-specific second mention and the first mention of the entity is created (indicated by the arc in (2)). We create such links in order to facilitate the development of systems detecting coherence violations.

(2) *(a) Taylor's attorney could not be reached for comment Friday night.*
(b) Tony Taylor, 34, of Hampton, Va., has a plea-agreement hearing scheduled for 9a.m.

definite noun phrase without reference to previous mention (DNP-REF): Definite NPs are generally used in text to refer to entities that are already present in the discourse context. We mark definite NPs that violate this rule. For example, an NP such as *'**the** Adam Air Boeing'* should be used in a summary only if the plane has been mentioned previously.

indefinite noun phrase with reference to previous mention (INP+REF): Indefinite NPs are used to introduce new entities to the discourse. For example, the NP *'**an** Adam Air plane'* is not appropriate if the same plane has already been mentioned in the summary. In such cases, we assign this feature and create a link to the previous mention.

pronoun with missing antecedent (PRN-ANT): This feature is used if a pronoun does not have any syntactically possible antecedent in the summary, i.e., there is no antecedent that matches in number and gender. In example 3, which shows the beginning of a summary, the pronoun *he* does not have any possible antecedent.

(3) *The trial opens of 29 mostly Moroccan suspects charged with involvement in the Madrid train bomb attacks in March 2004, which killed 191 people and injured 1,824 in the worst terror strike Spain has ever known. ROME **He** is charged with 191 counts of murder ...*

pronoun with misleading antecedent (PRN+MISLA): Extractive summarization systems sometimes place sentences with pronouns such that they follow a sentence with a grammatically and semantically possible antecedent; this is not always the *correct* antecedent according to the source documents. We identify such cases by referring to both the model summaries created by humans and the source documents. This feature is marked on the pronoun, with a link created to the misleading antecedent.

(4) *Jeff George, curator at Sea Turtle Inc., a nonprofit turtle rescue group on South Padre Island, said 42 of the endangered juvenile green turtles were released Tuesday and 46 on Wednesday. They return to Mexican waters when they are mature and can grow to 500 pounds (225 kilograms). "Seventeen of* ***them*** *were arrested on board, one skipper and another 16 workers. ...*

The above example shows that the approach of simply glueing sentences together may result in coreference chains that do not make sense; it was not the turtles that were arrested. The model summaries produced by humans for this document collection contain the information that the pronoun *them* refers to fishermen who were arrested.

acronyms without explanations (ACR-EXPL): We mark acronyms that are not generally known and that are not explained in the summary. To come up with a list of well-known acronyms, we collected a list of potential acronyms from the source documents. We then asked a native speaker of English to identify all items on the list that could reasonably be expected to be familiar to North American readers of news articles, as this is the domain of the source documents, as well as the imagined target audience of the TAC summaries.

2.2. LQ violations on the clause level

The clause level allows for annotations on arbitrary spans, from single tokens to complete sentences. Several of these violation types mark relations between spans rather than features of individual spans.

incomplete sentence (INCOMPLSN): Incomplete sentences occur in many summaries; these are generally due to the use of sentence compression or to truncation in order not to exceed the maximum allowed summary length. We mark such sentences with this feature.

(5) *He also extended his sincere sympathies to the bereaved families and those injured in*

inclusion of datelines (INCLDATE): Example (6) shows a dateline as they often occur in the source documents. Their inclusion into a summary is not desired, and such clauses are marked with this feature.

(6) *GEORGETOWN, Pennsylvania 2006-10-05 16:53:53 UTC*

other ungrammatical form (OTHRUNGR): This feature catches all other possible cases of ungrammaticality, such as missing spaces, wrong punctuation or cases like example (7). We mark entire clauses or sentences with this feature; in the current release of the corpus, it is not intended as a token-level feature.

(7) *Police say shooter at Amish school told wife he molested years ago, dreamed of doing it again*

no semantic relatedness (NOSEMREL): We mark adjacent sentences that obviously do not have any semantic relation, i.e., the cases where a reader wonders what one sentence has to do with the other. In example (8), the two sentences are not placed in an order or context that would seem natural to a reader.

(8)
- *(a) It is popularly known as the 'pink city' because of the ochre-pink hue of its old buildings and crenellated city walls.*
- *(b) He said there was no justification for such killings.*

redundant information (REDUNINF): We create links between clauses that express the same information, as redundancy negatively affects the readability of summaries.

(9)
- *The suspect apparently called his wife from a cell phone shortly before the shooting began, saying he was "acting out in revenge for something that happened 20 years ago", Miller said.*
- *The gunman, a local truck driver Charles Roberts, was apparently acting in "revenge for an incident that happened to him 20 years ago.*

no discourse relation (NODISREL): Discourse connectives indicate relationships between spans of text. With extractive summarization in particular, it can happen that an explicit discourse connective (*'and'*, *'but'*, *'because'*,...) is no longer appropriate in the new context of the summary. In such cases, we create a link of this type between two adjacent sentences and additionally mark the connective.

(10)
- *(a) Taylor's attorney could not be reached for comment Friday night.*
- *(b)* ***And*** *the person who cooperates first gets the biggest reward*

3. Annotation process and corpus statistics

In this chapter, we describe the source of the data in our corpus, report the results of an inter-annotator agreement study and give an overview of the collected annotations.

violation type	counts A	counts B	matches	P(A:B) R(B:A)	R(A:B) P(B:A)	F_1
entity mention level						
FM-EXPL	36	26	22	61.1	84.6	70.9
SM+EXPL	6	4	4	66.7	100.0	80.0
DNP-REF	34	23	18	52.9	78.3	63.2
INP+REF	19	9	9	47.4	100.0	64.3
PRN+MISSA	18	9	8	44.4	88.9	59.3
PRN+MISLA⋆	1	2	1	100.0	50.0	66.7
ACR-EXPL⋆	1	1	1	100.0	100.0	100.0
total/macro-avg	**115**	**74**	**63**	**54.5**	**90.4**	**67.5**
clause level						
INCOMPLSN	43	44	41	95.3	93.2	94.3
INCLDATE	24	24	23	95.8	95.8	95.8
OTHRUNGR	29	29	23	76.7	74.2	75.4
REDUNINF	28	26	19	65.5	73.1	69.0
NOSEMREL⋆	4	1	0	0.0	0.0	0.0
NODISREL⋆	3	0	0	0.0	0.0	0.0
total/macro-avg	**131**	**124**	**106**	**83.3**	**84.1**	**83.6**

Table 1: **Inter-annotator agreement measured by precision and recall**. Lines marked with ⋆ are excluded from the averages due to low frequency.

3.1. Data

We use the MAE annotation tool (Stubbs, 2011) to create stand-off annotations for 1,935 extractive summaries created by 44 summarization systems for the TAC 2011 Guided Summarization Task (Owczarzak and Dang, 2011), as well as 50 summaries generated by the G-Flow summarization system (Christensen et al., 2013). We excluded the summaries generated by a set of six summarization systems, which are also part of the TAC data, as their approach to summarization is not extractive. The linguistic quality of current non-extractive summarization methods is much lower than that of extractive methods, as most of their output is more or less completely ungrammatical. Applying our annotation scheme to texts such as these was neither sensible nor possible.

The TAC challenge requires participating systems to create a 100-word summary from 10 news documents for each of the 44 topics. The G-Flow summaries have been created for 50 document clusters, each containing about 10 documents, from the DUC 2004 Summarization Task. Our primary annotator marked LQ violations for the entire set of summaries, identifying 5,752 instances of violations of linguistic quality.

3.2. LQVSumm: Corpus statistics

All 1,935 extractive TAC summaries and 50 G-Flow summaries have been annotated by our primary annotator (A in the following section). Table 2 shows the annotation counts for the entity mention level and the clause level. The most frequent violation types are definite NPs without previous references to the same entity (DNP-REF) and first mentions of an entity that lack clear referents (FM-EXPL). Pronouns with missing or misleading antecedents occur less often, as some systems already have components that detect such cases.

On the clause level, incomplete sentences are the most frequent violation type. Most systems make use of the full 100 words in the TAC challenge, even if that means truncating sentences in order to raise their content scores. This study, however, shows that ending summaries with incomplete sentences decreases the quality of the summary in terms of readability.

On average, each TAC summary contains 2.96 violations, while each G-Flow summary contains only 0.5 violations, indicating that G-Flow is indeed successful at creating coherent summaries.

3.3. Inter-annotator agreement

In order to test the reliability of the annotation scheme, our primary annotator (A) trained a paid undergraduate student of computational linguistics (B) for the annotation task, using 20 summaries as training material. Then, each annotator independently marked a set of 100 summaries (95 TAC / 5 G-Flow). As the annotation task consists of both a detection and a labeling task, we report agreement as precision (P) and recall (R) when treating A as the gold standard and B as a 'system' (notated as P(B:A), R(B:A)) and vice versa. Note that P(A:B)=R(B:A) and P(B:A)=R(A:B).

In contrast to a single metric such as Cohen's κ, the precision-recall based analysis immediately shows whether one annotator marks more instances of a given violation type than the other, and how many annotations are marked by both. We count both exact span matches and overlapping spans as matches, provided they are labeled with the same violation type.

Table 1 shows agreement for violations of coherence on the **entity mention** level. Most of the annotations created by B match annotations of A, resulting in high values for P(B:A) and R(A:B). Annotator A creates about twice as

many annotations of violations. We attribute part of this to her greater experience in the annotation task. Agreement is high for individual annotations as well as for links related to particular LQ violations. For the violations types SM+EXPL, INP+REF, and PRN+MISLA, B created 18 links to first or previous mentions. 16 of these match links created by A. Overall, this shows that the degree of subjectivity of the annotation task on the entity level is manageable, but that it requires a high degree of diligence when trying to create a more or less complete annotation of all violations.

The agreement for annotations on the **clause level** is also listed in Table 1. In this case, more than 83% of each of A's and B's annotations can be found in the other's annotations as well. Violations on the clause level seem to be easier to detect than those on the entity level.

4. Analysis / Modeling

In this section we investigate the relationship between the number of annotated violation types in our corpus and the manually assigned evaluation scores from the TAC challenge. We do this in two ways: first, on the summary level, we measure the **correlation** between manually assigned scores and the number of LQVSumm annotations. Second, on the system level, we compare **rankings** of summarization systems according to either the average score assigned to their summaries or the average number of LQ violation annotations.

4.1. *Pyramid*, *Readability* and *Responsiveness* scores

During the TAC 2011 challenge (Owczarzak and Dang, 2011), in addition to the *Pyramid* scores, which reflect content coverage, each summary was manually evaluated with respect to *Readability* and *Responsiveness*. *Readability* focuses on LQ: judges were asked to judge how fluent and readable each summary is independently of whether it contains any relevant information.[4] The score is intended to reflect grammaticality, non-redundancy, referential clarity (it should be clear who the noun phrases in the summary refer to), focus, structure and coherence all at once. A more elaborate description of these factors was given in the context of the Document Understanding Conferences (DUC).[5] *Responsiveness* judges both content coverage and LQ. Both are marked on a 5-point scale: (1) very poor, (2) poor, (3) barely acceptable, (4) good, (5) very good.

4.2. Summary-level correlation with *Readability* scores

We sum up the number of annotations per violation type per summary, and compute Pearson's correlation coefficient r between these sums and the *Readability* scores assigned to the respective summaries. As the number of violations is expected to be inversely proportional to the *Readability* score (the more violations, the lower the *Readability* score), and most violation types occur in only a few documents, we mostly observe weak negative relationships. Table 2 shows the correlation coefficients by violation type for all three scores: *Readability*, *Responsiveness*, and *Pyramid*.

Entity-level violations. On the entity level, only definite NPs without previous references to the same entity (DNP-REF) and pronouns that lack antecedents (PRN+MISSA) have an effect on *Readability*, and both of these also exhibit negative correlations with *Pyramid* and *Responsiveness*. It is interesting to note that while indefinite NPs with previous references to the same entity (INP+REF) are not correlated to *Readability*, they are relatively strongly positively correlated to *Pyramid*. Sentences containing indefinite NPs often contain important information, as they introduce a new entity to the discourse, and hence their occurrence in a summary seems to lead to good content coverage.

When summing over the two violation types with the highest absolute correlation values (DNP-REF and PRN-MISSA) per summary and correlating these sums to the evaluation scores, we observe a stronger relationship for all of the three scores. Finally, we sum over all entity level violations and compute the correlation with the three scores. The strongest negative correlation is observed for *Readability*. There is no correlation between these sums and the *Pyramid* scores, and weak correlation to *Responsiveness*.

Clause-level violations. On the clause level, all violation types except for the infrequent NODISREL show negative correlations to *Readability*. Redundant information (REDUNDINF) is positively correlated to *Pyramid*. We assume that systems picking redundant sentences do so because these sentences are highly ranked with regard to content coverage. Summing all clause level violations and correlating to *Readability* results in a moderate negative relationship.

All violations. Summing over all violation types per summary and correlating to the evaluation scores again shows a moderate negative relationship to *Readability*, no correlation to *Pyramid*, and a weak negative relationship with *Responsiveness*. This shows that the *Pyramid* score is not influenced by the types of LQ violations we annotate, while the *Readability* and *Responsiveness* scores are. This result is expected, as the *Pyramid* score is not supposed to reflect linguistic quality, but *Responsiveness* should.

If we use all clause level violation types and only the two entity level violation types with the highest absolute correlation coefficients, we observe even stronger correlations, but the tendencies regarding the different evaluation scores stay the same.

The correlation coefficients as presented in this section show that when considering the entire collection of summaries, the occurrences of most violation types annotated in LQVSumm have an influence on the manually assigned *Readability* scores. However, some violation types achieve only low values for Pearson's r because they occur extremely infrequently. Nevertheless, they could have a

[4] http://www.nist.gov/tac/2011/Summarization/Guided-Summ.2011.guidelines.html

[5] http://www-nlpir.nist.gov/projects/duc/duc2006/quality-questions.txt

corpus	G-Flow 50 documents		TAC 1,935 document				
violation type	count	avg/doc	count	avg/doc	Pearson's *r*		
					Readability	Pyramid	Respons.
entity level violations							
DNP-REF	3	0.06	958	0.50	**-0.122**	**-0.166**	**-0.133**
FM-EXPL	6	0.12	792	0.41	0.006	-0.050	**-0.066**
INP+REF	1	0.02	430	0.22	-0.052	**0.235**	**0.109**
PRN+MISSA	2	0.04	361	0.19	**-0.191**	**-0.140**	**-0.156**
SM+EXPL	1	0.02	162	0.08	0.020	**0.089**	0.045
PRN+MISLA	0	0.00	27	0.01	**-0.065**	**-0.073**	**-0.089**
ACR-EXPL	3	0.04	11	0.01	-0.038	-0.056	-0.006
sum(DNP-REF, PRN+MISSA)	5	0.1	1319	0.68	**-0.204**	**-0.208**	**-0.192**
sum(entity level violations)	15	0.03	2741	1.42	**-0.167**	-0.074	**-0.127**
clause level violations							
INCOMPLSN	0	0.00	1,044	0.54	**-0.210**	0.000	-0.029
OTHRUNGR	3	0.06	793	0.41	**-0.180**	0.007	-0.016
INCLDATE	3	0.06	412	0.21	**-0.090**	0.039	0.051
REDUNDINF	3	0.06	504	0.26	**-0.160**	**0.156**	**0.077**
NOSEMREL	0	0.00	142	0.07	**-0.148**	**-0.102**	**-0.132**
NODISREL	1	0.02	91	0.05	-0.025	**-0.081**	**-0.062**
misleading discourse connectives⋆	1	0.02	114	0.06	-	-	-
sum(clause level violations)	10	0.2	2,986	1.54	**-0.325**	0.041	-0.016
sum(clause level violations, DNP-REF, PRN+MISSA)	15	0.3	4,305	2.22	**-0.385**	**-0.084**	**-0.122**
sum(all violations)	25	0.5	5,727	2.96	**-0.356**	-0.022	**-0.101**

Table 2: **Counts of annotations of coherence violations** marked in the TAC and G-Flow data sets and **Pearson's correlation coefficients** (r) of the number of violations per document and the manually assigned scores from TAC 2011. Bold numbers indicate significance at $p < 0.01$. ⋆ excluded from the averages.

strong influence on a judge's decision to give a low *Readability* score if they occur in a summary. In the next section, we address this question.

4.3. Summary-level linear regression model for predicting *Readability*

Another way of estimating the relative impact of occurrences of the violation types on the manually assigned *Readability* scores is to inspect their weights when used as features in a linear model to predict *Readability* scores. To do this, we fit a linear model using the `lm` function of `R`, using the number of occurrences of each violation type per summary as features and predicting the *Readability* score. We use the entire data set in order to fit this model. Table 3 shows the coefficients for the linear model. Except for SM+EXPL, all violation types have negative coefficients, i.e., their occurrence in a summary leads to a decrease of the predicted score. The clause level violation types identified as contributing to a decreased *Readability* score in the previous section all get significant negative coefficients in this experiment as well. In addition, it is interesting to note that some entity level violation types that are not correlated to the *Readability* scores *do* have an effect on the score assigned to a particular summary. These types are acronyms without explanations (ACR-EXPL) and pronouns with misleading antecedents (PRN+MISLA).

The only violation types that do not achieve meaningful correlations to the summary-level *Readability* score or only coefficients on small magnitude in the linear model are indefinite NPs with previous references (INP+REF), no discourse relation (NODISREL), first mention without explanation (FM-EXPL) and subsequent mention with explanation (SM+EXPL). We conclude that of our LQV types, all except these four clearly influence judges when assigning *Readability* scores. The four types either occur too infrequently in the data set to have an effect in the statistical evaluation as presented here, or do not hurt the perceived readability of a text to a extent sufficient to cause judges to give lower scores.

Feature	Weight	Feature	Weight
Intercept	**3.407**	DNP-REF	**-0.157**
ACR-EXPL	-0.361	OTHRUNGR	**-0.155**
PRN+MISLA	**-0.355**	INCLDATE	**-0.151**
INCOMPLSN	**-0.275**	INP+REF	-0.067
NOSEMREL	**-0.262**	NODISREL	-0.046
REDUNDINF	**-0.259**	FM-EXPL	-0.023
PRN+MISSA	**-0.236**	SM+EXPL	0.038

Table 3: **Linear Model** for the full training set. Bold numbers indicate significance at $p < 0.01$.

Ranking using LQVSumm annotations						TAC 2011 Ranking					
all violations		entity level violations		clause level violations		Readability		Pyramid		Respons.	
*ID**	*score*	*ID*	*score*	*ID*	*score*	*ID*	*score*	*ID*	*score*	*ID*	*score*
G**	0.5	G	0.3	G	0.2						
21	1.30	**1**	0.34	**16**	0.23	**32**	3.75	**22**	0.47	**25**	3.16
32	1.30	**2**	0.75	**37**	0.25	**21**	3.52	**43**	0.47	**22**	3.14
1	1.34	**9**	0.80	**22**	0.43	**48**	3.50	**17**	0.46	**13**	3.11
37	1.43	**21**	0.84	**21**	0.45	**37**	3.45	**4**	0.45	**17**	3.09
48	1.55	**32**	0.84	**32**	0.45	**22**	3.43	**28**	0.44	**21**	3.09
2	1.75	**10**	0.89	**34**	0.45	**25**	3.34	**24**	0.44	**32**	3.09
					...						
31	4.27	**4**	1.93	**6**	2.70	**33**	2.59	**45**	0.32	**34**	2.50
33	4.32	**39**	1.98	**10**	2.75	**5**	2.57	**8**	0.31	**29**	2.43
18	4.34	**40**	2.00	**45**	2.89	**23**	2.50	**6**	0.31	**8**	2.36
23	5.30	**5**	2.07	**31**	3.30	**31**	2.50	**14**	0.31	**23**	2.34
13	5.45	**14**	2.11	**33**	3.41	**6**	2.32	**23**	0.30	**6**	2.34
45	5.57	**45**	2.68	**13**	4.23	**45**	2.27	**1**	0.30	**45**	2.27
7	5.77	**23**	3.07	**7**	4.63	**11**	2.09	**11**	0.28	**11**	2.23

Table 4: **Ranking** of systems of the TAC 2011 Guided Summarization task (initial summaries) according to different metrics. *ID = summarization system ID. **G = G-Flow.

4.4. System-level rankings

In the TAC 2011 Guided Summarization Task (Owczarzak and Dang, 2011), 50 systems participated and were evaluated according to *Readability*, content coverage (*Pyramid*) and *Responsiveness*. Table 4 shows the ranking of the 44 systems whose summaries are annotated in our corpus according to these scores. The system IDs are the ones used in the TAC challenge, and we only rank the 'initial' summaries.[6] In addition, we rank the systems by the number of violations of linguistic quality. Comparing these rankings allows us to investigate the type of problems that a system has regarding the readability of its summaries. The following list, naming some conclusions that can be drawn by inspecting the ranking, is not exhaustive:

- Out of the top-5 ranking systems both for *all violations* and *Readability*, 4 overlap. This shows that systems whose summaries were marked with only few violations in our corpus also achieved the best *Readability* scores in the TAC 2011 challenge.
- Comparing the lower end of the rankings, we observe more variation, although systems stay approximately in the same region.
- System 1 is a baseline that produces a 'summary' by simply extracting the first 100 words of one source document. It is the system with the fewest entity level violations, but it has one clause level violation (incomplete sentence) per summary. This is also reflected in the *Readability* score, and shows how puzzled readers are by summaries containing such sentences. This likely increases the reading time of summaries, which is counterproductive concerning the aim of summarization.
- The top systems for the content-based score *Pyramid* rank in the middle region for *all violations* and *Readability* (except for system 22 which has rank 5 in the *Readability* ranking). This suggests that no system yet adequately addresses both content coverage and linguistic quality. Our fine-grained analysis highlights the types of problems reflected in the output from individual systems; for example, the best system with respect to content coverage (system 22) has few violations on the clause level, but many on the entity level.

4.5. Correlation of system-level rankings

Table 5 shows the correlation between the system-level scores of systems according to our corpus-based evaluation (we compute the average number of violations per summary as also shown in Table 4) and the *Readability* scores. We report the Pearson's r, Spearman's ρ and Kendall's τ for a state-of-the-art system for automatically creating readability scores, DICOMER by Lin et al. (2012). DICOMER uses features based on a PDTB-style discourse parser and is trained on data from the TAC 2009 and 2010 multidocument summarization challenges. Lin et al. (2012) use the logarithm of their predicted scores when computing r, as r reflects the linear relationship between two variables. Inspection of our data shows that the relationship between the number of violations per summary and its *Readability* score is of a logarithmic shape, hence, we also take the logarithm of the number of violations when computing r.

We compare the magnitude of the correlations, as DICOMER directly predicts the *Readability* scores while the number of violations per document is expected to be lower for systems with high *Readability* ratings. DICOMER evaluates all 50 systems, while our results are reported for the

[6] There was a second task to create 'update' summaries, which we don't address yet.

44 systems whose summaries are annotated in our project (see section 3.1), and the numbers are therefore not directly comparable. However, they differ enough to suggest the following conclusions.

DICOMER performs better than our prediction in terms of Pearson's r, but this is not surprising given that DICOMER has been trained on *Readability* scores, while our system consists of a simple heuristic. Spearman's ρ and Kendall's τ, on the other hand, only evaluate the relative ranking of systems, respectively taking the differences of the scores into account or not. Our annotation-based method actually creates a ranking of summarizers that is closer than DICOMER's to the one induced by the *Readability* scores. This is a very promising result as it shows that there is headroom for the development of automatic methods for the evaluation of linguistic quality, and that our corpus of annotations of linguistic quality violations is a valuable resource for such research.

Method	r	ρ	τ
DICOMER (Lin et al., 2012)	**0.867**	0.712	0.535
LQVSumm: $\sum$(# violations)	-0.82	**-0.858**	**-0.713**

Table 5: **Correlation** (**Pearson's** r, **Spearman's** ρ, **Kendall's** τ) of system-level scores with *Readability* scores for the TAC 2011 Guided Summarization Task.

5. Related work

In this section, we briefly review related work in the area of the evaluation of the linguistic quality of automatically created summaries. As previously mentioned, the definition of *Readability* (see section 4.1) was first introduced in the context of the shared tasks on summarization of the Document Understanding Conferences. The same definition is used in the TAC challenges. For the first time, in 2011, the shared task on summarization organized by TAC included the automatic judgment of readability in the task on Automatically Evaluating Summaries of Peers (AESOP). The data annotated in our work originates from the data released in the context of this task. The top-performing systems (de Oliveira, 2011; Conroy et al., 2011; Giannakopoulos and Karkaletsis, 2011; Kumar et al., 2011) use n-gram based matching techniques, comparing the summaries to source texts or model summaries.

Lapata and Barzilay (2005) and Barzilay and Lapata (2008) address the problem of the automatic evaluation of the local coherence, i.e. sentence-to-sentence transitions, of summaries. Their probabilistic approach models both entity coherence and lexical cohesion. Guinaudeau and Strube (2013) propose a graph-based model for the same task.

Pitler and Nenkova (2008) collect readability judgments for texts from the Penn Discourse TreeBank (Prasad et al., 2008) and find that syntactic, semantic and discourse-based features are good predictors of readability. Using these insights, Pitler et al. (2010) train a linguistic quality model using various linguistic features aimed at capturing the coherence and fluency of a summary. Lin et al. (2012) propose an automatic method to evaluate summary readability by incorporating features extracted by a Penn Discourse Treebank-style parser.

Further research on judging the readability of texts is done for non-automatically generated texts. For example, Van Oosten and Hoste (2011) create a corpus of readability judgments by having expert readers as well as non-experts score and rank pairs of texts for their readability. They intentionally do not give a definition of readability in order to model readability as generally as possible. Recently, Persing and Ng (2013) present a corpus of 830 essays written by learners of the English language annotated with clarity scores. Their work is similar to ours in that they also aim to identify particular classes of errors.

6. Conclusion and Future Work

Research on automatic summarization is currently changing its focus from content selection to creating coherent summaries, or summaries of good linguistic quality (Nenkova and McKeown, 2011; Christensen et al., 2013). The evaluation of linguistic quality is costly, as it is mostly done manually to date. The research community is in need of automatic methods for evaluating the linguistic quality of summaries. We contribute to their development by publishing a corpus of 1,985 automatically created summaries annotated with violations of linguistic quality.

We have shown that there are relationships between the types of violations defined in our annotation scheme and the intuitively assigned *Readability* scores of the TAC 2011 challenge. The annotations also reveal strengths and weaknesses of the summarization systems with regard to particular violations of linguistic quality, and immediately point to possibilities for improvement of the respective systems.

Future work comprises the creation of methods and tools that detect the violations in text, automatically evaluating the linguistic quality of summaries not only by outputting a numeric score that correlates well with *Readability* scores, but also offering a diagnostic instrument.

Also, summaries from the TAC 2011 Update Summarization Task are not included in the corpus so far, as their annotation requires further study of the texts and possibly additions to the annotation scheme. As the creation of summaries updating a reader on a particular topic has a great relevance, the extension of our annotation scheme to this type of summary is an important next step.

Finally, it may be interesting in future work to investigate whether this approach can be modified to suit summaries automatically created by systems aiming at abstractive summarization techniques.

Acknowledgments

This research was supported in part by the MMCI Cluster of Excellence. The first author is supported by an IBM PhD Fellowship, and the second author by an Erasmus Mundus scholarship of the LCT program. We thank Manfred Pinkal for his advice concerning the development of the annotation scheme, and Jonathan Oberländer and Prashanth N Rao for participating in the inter-annotator agreement study.

7. References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards Coherent Multi-Document Summarization. In *Proceedings of NAACL-HLT*, pages 1163–1173.

John M Conroy, Judith D Schlesinger, Jeff Kubina, Peter A Rankel, and Dianne P O'Leary. 2011. CLASSY 2011 at TAC: Guided and multi-lingual summaries and evaluation metrics. In *Proceedings of the Text Analysis Conference*.

Paulo CF de Oliveira. 2011. CatolicaSC at TAC 2011. In *Proceedings of the Text Analysis Conference (TAC 2011), Gaithersburg, Maryland, USA, November*.

George Giannakopoulos and Vangelis Karkaletsis. 2011. AutoSummENG and MeMoG in evaluating guided summaries.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 93–103.

Niraj Kumar, Kannan Srinathan, and Vasudeva Varma. 2011. Using unsupervised system with least linguistic features for tac-aesop task. In *Fourth Text Analysis Conference (TAC 2011)*.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *IJCAI*, volume 5, pages 1085–1090.

Ziheng Lin, Chang Liu, Hwee Tou Ng, and Min-Yen Kan. 2012. Combining coherence models and machine translation evaluation metrics for summarization evaluation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 1006–1014. Association for Computational Linguistics.

Ani Nenkova and Kathleen Rose McKeown. 2011. *Automatic summarization*. Now Publishers Inc.

Karolina Owczarzak and Hoa Trang Dang. 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*.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. pages 260–269, August.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195. Association for Computational Linguistics.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 544–554. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

Amber Stubbs. 2011. MAE and MAI: lightweight annotation and adjudication tools. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 129–133. Association for Computational Linguistics.

Philip Van Oosten and Véronique Hoste. 2011. Readability annotation: Replacing the expert by the crowd. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 120–129. Association for Computational Linguistics.