Saarland University

Department of Computational Linguistics

Master Thesis

Annotation of Factors of Linguistic Quality for Multi-Document Summarization

Author:

Marina Valeeva

Matriculation: 2546520

Supervisors:

Prof. Dr. Manfred Pinkal

Dr. Alexis Palmer

Annemarie S. Friedrich

September 30th, 2013

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Saarbrücken, September 30th, 2013 Marina Valeeva

Abstract

The increasing availability of automatically generated summaries has prompted intensive research in the area of automatic text summarization evaluation within the Natural Language Processing community. The development of automatic summarization systems has mainly focused on improving content selection, which has led to a focus on automatic methods for evaluation of the content of automatically generated summaries. The problem of evaluating the linguistic quality of summarization system outputs has been addressed only in recent years.

Linguistic quality of a summary is associated with its readability, which includes various linguistic factors that determine the overall quality of a summary. This project investigates and defines a set of potentially important factors of linguistic quality for further manual annotation, with purpose of revealing linguistic quality violations that occur frequently in machine-produced summaries. The results of this analysis motivate a scheme and process for annotation of such violations in automatically produced summaries. The annotation scheme that has been developed for this task and further employed in the process of annotation includes five annotation tags: entity mention, clause, relation between entities, relation between clauses and misleading discourse connective.

This project's outcome is a corpus of summaries annotated for the identified factors of linguistic quality. The corpus contains 1935 annotated extractive summaries from the TAC dataset and 50 annotated extractive summaries from the G-Flow dataset. Statistical analysis of both annotated datasets has been carried out and evaluated. The results actually show that the G-Flow system produces summaries with fewer violations on average than the TAC systems. This is the most we can say, especially given the relatively small G-Flow dataset.

For producing the corpus a detailed set of annotation guidelines containing the explanation of annotation tags, the classification into attribute types followed by examples, has been developed. Inter-annotator agreement has been measured to verify the reliability of the guidelines and returns high agreement rates for most annotation tags.

Acknowledgements

I would like to express profound gratitude to the following people for their invaluable contribution and assistance in conducting my thesis work:

To Prof. Dr. Manfred Pinkal for his guidance, coaching and extremely useful comments;

To Dr. Alexis Palmer and Annemarie S. Friedrich for being kind, supportive and ready to share their personal experience and great ideas always and in any situation;

To Jonathan Oberlaender and Prashanth N. Rao for their great job in completing the manual annotation task.

And also,

To my family for enormous emotional support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Advances of civilization accelerated the growth of knowledge and availability of information to people all over the world. Books and news publications coming out both in paper and electronic formats increased the volume of data available as online texts. Automatic text summarization is of enormous service to readers by condensing large or multiple documents available online in different natural languages and forming shorter summaries while still preserving core information.

Automatic text summarization has become a valuable aspect of our everyday life. Automatic text summarization not only contributes to people's efficient processing of various facts, but it also allows us to save time while gaining knowledge. To automatically generate a good summary of a set of documents is a process where a summarization system, on one hand, has to extract the most relevant information from the original text, and on the other has to consider the fluency of the output text. So far automatic text summarization has not yet reached a level of quality which is comparable to manually created summaries.

In the course of this project the linguistic quality of automatically generated summaries has been analyzed and a number of the most common violations of readability have been identified. In order to understand how such violations of linguistic quality influence the readability of summaries, two datasets containing extractive summaries of newswire text have been annotated according to the scheme we designed for this task. The annotation guidelines have been designed to annotate the phenomena that have been violated for entity mention, clauses and discourse connectives. Therefore, the main objective of this work has been fulfilled by constructing a manually annotated corpus for such violations, which includes 1935 extractive summaries from the TAC dataset and 50 exctractive summaries from the G-Flow dataset. The annotation results of both datasets have been compared and evaluated.

Evaluation of summaries has become a special domain of research in recent years where not only the content coverage plays an important role in overall quality of a final summary, but also its other properties such as coherence and readability of the text. Readability is associated with a summary's fluency and is based on a number of aspects of linguistic quality.

In the TAC 2011 Guided Summarization Task, readability is assessed by humans that manually rate summaries based on particular set of linguistic criteria. The advantage of this manual rating is that it is rather fast and cheap and does not require a gold standard, or any annotation from an assessor to evaluate the summary's linguistic quality and come up with a readability score. The disadvantage of this manual metric is that there is no inter-human agreement or reproducibility (Nenkova and McKeown, 2011). Both these factors make humans assign different readability scores to one and the same summary, additionally one and the same assessor may perceive information differently at different periods of time and therefore, assign different readability scores to the same summary.

This project aims to provide a detailed analysis of linguistic quality of summaries with the purpose of detecting those types of violations of linguistic quality that influence readability scores. I focus on the manual analysis of linguistic quality of automatically generated summaries in order to analyze and identify the aspects of linguistic quality that might have an effect on the overall readability of the summary. The hypothesis being tested in this work is that summaries scoring low on readability will have more violations of certain types than those that have been scored high. Therefore, I intend to identify the dependencies between certain violations of linguistic quality and readability scores and determine those violations that influence the readability score.

In this work, I have manually analyzed readability of summaries generated by various summarization systems. The results of this analysis showed that extractive summaries which are generally produced by simple sentence extraction from the original document still suffer from some sentences being extracted out of context that leads to the problem of semantic relatedness or dangling anaphors. I have detected a set of violation types that might lead to coherence problems in summaries. These violations have been organized into annotation classes for each major category of linguistic quality violation to form the annotation scheme.

Finally, I use the annotation scheme to produce an annotated corpus with the aim to contribute to the development of automatic linguistic quality measurements in multi-document summarization.

After providing the motivation for this project and addressing the challenges involved in evaluating linguistic quality of machine-generated summaries in Chapter 1, Chapter 2 on general background of automatic summarization describes various types of summarization and also introduces a range of approaches to automatic summarization. Chapter 3 presents related work for summarization evaluation and the contributions of research on the task of evaluating linguistic quality in multi-document summarization. Chapter 4 describes the two datasets used for manual annotation and further creation of the corpus of annotated summaries. Chapter 5 contains a discussion of the analysis of the summaries and describes the annotation scheme that has been developed as a result of this analysis, the annotation process and the annotation tool used for the manual annotation. Chapter 6 provides the statistical corpus data analysis and the comparison of two datasets, TAC and G-Flow. Chapter 7 concludes this work by providing a discussion of the results of the annotation process and the outlook on future work in this area of research.

Chapter 2

Automatic Summarization: General Background

Most of existing research in automatic text summarization resulted in the development of various summarization approaches, methods and algorithms. The availability and variability of such approaches is important as they allow for comparison studies and continuous improvement with the purpose of generating content-rich and coherent summaries.

In this section, previous approaches to automatic summarization are described. I pay special attention to approaches that are reflective of the summarization systems that participated in the TAC 2011 Guided Summarization Task because these systems generated the summaries which I use to produce an annotated corpus for various violations of linguistic quality.

2.1 Extractive vs. Abstractive Summarization

Automatic text summarization can be of two types: abstractive and extractive. Previous research mainly focuses on extractive summarization wherein summaries are created by selecting the most salient sentences in a document or a set of documents and putting them together. Extractive summarization is considered to be an easier and more practical way of summarizing texts than abstractive summarization, which typically creates summaries by generating new text based on the desired content (Hovy, 2005). The extractive summarization approaches mainly focus on content selection. Even if the content is well chosen, linguistic quality of output summaries that ensures that a summary is coherent and easy to read and comprehend still remains a daunting challenge.

Automatic summarization can be applied either to a single document or to multiple documents. Summarizing a single document is a hard task. Summarizing a set of documents related to one topic is even harder, posing additional challenges. In multi-document summarization where multiple sources of information overlap or supplement each other, extracts can often be incoherent, due to repeated or omitted material, dangling references, etc (Hovy, 2005). In this case a simple concatenation approach to extraction would not work. The task then will be not only identifying the most important parts of a document but also capturing new information and avoiding redundancy. Taking into account linguistic quality in the process of evaluation of summaries ensures that a final summary is not only complete content-wise, but also readable.

2.2 Summarization Approaches

2.2.1 Earlier Approaches

There has been a long history of research in text summarization. Most studies on text summarization today rely mainly on sentence extraction to produce a summary. A great number of various approaches were developed for creating “good” extractive summarization systems and some of them are still followed today as foundations for text summarization. In this section just a few of these earlier approaches are discussed.

One of the most prominent works which made a valuable contribution to the field of text summarization is Luhn (1958)’s research that suggested extracting salient sentences from the text by using features like *word frequency*. According to Luhn (1958) extractive summarization systems were based on retrieving the most frequent words and sentences from the original document where he would also put together a list of words sorted by their frequency where a certain index would show how important a particular word is. Within a sentence, a significance factor demonstrated how often significant words occurred and what is the distance between these words by indicating the number of non-significant words that take place in between. Therefore, all sentences were assigned a certain significance factor and only the sentences with the highest significance factor were selected to create a summary.

Further advances in text summarization were achieved by Baxendale (1958) and Edmundson (1969). Baxendale (1958) proposed the *sentence position* within the text to be an important feature for defining salient part of the document. And Edmundson (1969) suggested to extend features such as *word frequency* and *sentence position* with two new features *cue words* and *title words*. *Cue words* and *title words* cause greater weight to be assigned to sentences containing cue words like ”significant” or ”impossible” or words that appear in the title of a document. According to Baxendale (1958), the position of the sentence in a document gives a good clue to the importance of the sentence in a document.

Baxendale (1958) proposed a fairly precise way to select a topic sentence from a document by selecting sentences which appear at the very beginning and at the very end of the document and/or each paragraph (Jurafsky and Martin, 2000). Lin and Hovy (1997) found the feature of sentence position also important. Instead of combining a number of features they concentrated only on one single feature insisting on a so-called *position method* according to which sentences of greater topic centrality take place in certain predictable positions such as titles of the document. However, this approach significantly depends on the text genre and the subject domain and therefore cannot be fully relied on (Lin and Hovy, 1997).

In summary, most earlier works on extractive summarization are dependent on ranking sentences based on the computed scores when one or more features are taken into account: term frequency, sentence position, cue words, title words and then selecting n top ranked sentences (Luhn, 1958; Baxendale, 1958; Edmundson, 1969; Lin and Hovy, 1997).

2.2.2 Multi-Document Summarization as Sentence Selection

Since 2007 the Text Analysis Conference (TAC)[1] represents a series of annual NIST-led evaluation workshops that support research within the Natural Language Processing community. The Guided Summarization Task has been one of the main tasks of TAC where a summary should cover the aspect-oriented information for the categories. The Guided Summarization Task encourages the systems that participate in this task to have a deeper analysis of the content of documents instead of just employing the word frequency to extract the content. This task is aspect-guided as there are just aspects that cover important information that help to understand the specific content of a document set. Aspects are predefined for each category. The full list of such categories and aspects is provided in Chapter 4.

For example, the Accidents and Natural Disasters category has the following aspects:

- WHAT: what happened
- WHEN: date, time, other temporal placement markers
- WHERE: physical location
- WHY: reasons for accident
- WHO AFFECTED: casualties (death, injury), or individuals otherwise negatively affected by the accident/disaster

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

- DAMAGES: damages caused by the accident
- COUNTERMEASURES: countermeasures, rescue efforts, prevention efforts, other reactions to the accident

Most multi-document summarization approaches that participated in the TAC 2011 Guided Summarization Task[2] aimed to generate summaries in an extractive way, which means that the summarization systems will extract whole or partial sentences from the original documents and then put them together to form the summary. Original documents are the documents that are used as inputs for summarization systems to produce summaries out of them. This task involves the participant systems to make a deeper linguistic (semantic) analysis while focusing on aspect coverage, so that content selection for the summary is category oriented and is based on its relevance to the main topic of the original document set.

Feature based approach: There are various types of approaches for document summarization. One of the most frequently used methods in extractive type of summarization is the feature based approach. According to this approach different types of features are determined and based on these features the relevance and importance of the sentences are identified and then such sentences are selected for the summary. General extractive multi-document summarization systems are based on word frequency, where the words with the highest frequency (except for stop words) are more likely to be relevant to the topic of the document set and thus, picked by the system to generate the summary. The motivation for using the word frequency is that important words tend to appear in the document multiple times. The approach for sentence selection that has been proposed by Mason and Charniak (2011) is a word frequency-based approach but with a slight modification where a negative weight is assigned to sentences that contain document-specific words. The reason for penalizing the sentences that contain too specific content is that good summaries should contain information that is relevant to the main ideas of the entire multi-document set, and, therefore, should not contain information that is too specific to any single document in the set (Mason and Charniak, 2011).

The query-focused summarization system CLASSY by Conroy et al. (2011) introduces the query term selection component that impacts sentence scoring and also computes two sets of features - content features and those that aim to improve linguistic quality. Conroy et al. (2011) use a rich set of query terms that includes the title words, category terms and also aspect terms. Title words is the commonly used feature type for the summarization task because the co-occurrence of such words in a sentence with those in the document's title indicates that the sentence is highly relevant to the document

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

and therefore, worth selecting for the summary. Other important types of features are sentence position and sentence length. The sentence position helps to identify the location of important information in a document. The beginning of the article often tends to contain a lot of important sentences. The sentence length, in its turn, assists to eliminate sentences that are too short to be included into the summary as they cover less important information. Too long sentences are not selected to form a summary either.

Not all features can be treated with the same level of importance. Therefore, most summarization approaches incorporate a combination of different features as it tends to improve the results. Ng et al. (2011) introduce a robust extractive multi-document summarizer SWING based on the combination of generic features, such as sentence length, sentence position, document frequency and category-specific features that computes the so-called *category-specific importance* (CSI) of sentences. SWING also makes use of named entity information at the topical level as named entities are known to be indicators of important information. In their SemQuest, a question-answering system, Barrera et al. (2011) propose a ranking-based information extraction method where they also exploit the use of named entities. Sentences are sorted, ranked with regards to the scores assigned to them and then selected for the summary.

In their extractive summarization model, Varma et al. (2011) implement knowledge based measures (document frequency, sentence position and prepositional importance) using Wikipedia articles for extracting important sentences from documents and topic modeling with Latent Dirichlet Allocation (LDA) for scoring sentences. Among the top ranked sentences, the sentences that have the maximum aspect scores are selected for the summary. Liu et al. (2011) introduce an extractive multi-document summarization system based on hierarchical topic model of hierarchical LDA (hLDA). hLDA is a representative generative probabilistic model, which not only tries to reveal latent topics from a large amount of discrete text data, but also can organize these topics into a hierarchy for achieving a deeper semantic analysis. The hLDA model used by Liu et al. (2011) aims to discover topics in the collection of sentences and then organize these topics into a hierarchy. This model also combines some traditional features such as the similarity of title sentences, the number of named entities per sentence, the word frequency and the keyword coverage. Liu et al. (2011) calculate the similarity between sentences in order to remove redundant information and also use the sentence compression technique to make a summary coherent and more readable.

Li et al. (2011a) present two different extractive summarization approaches based on sentence selection. The first approach scores sentences by linearly combining manifold ranking scores and scores based on other features (for instance, sentence position and named entity coverage) and the second approach uses the extended version of the traditional Integer Linear Programming (ILP) - Tolerated ILP where a certain degree of redundancy is accepted, meaning that concepts can be selected multiple times into the summary.

Klassen (2011) implements a basic feature-based extractive summarization system that employs Log Likelihood Ratio (LLR) to calculate signature terms. The features used by Klassen (2011) are based on POS tags, named entity and dependency information. Contrary to the previously mentioned feature-based approaches, Kumar et al. (2011) make use of an unsupervised system with no linguistic features, where the local importance of words is used for calculating the importance of topics.

Graph based approach: Hu and Ji (2011) introduce their WHUSUM query-focused summarization system based on aspect-related sentence selection and graph-based sentence ranking algorithm to facilitate the extraction of aspect-related sentences with informativeness and diversity. To guide the summarization process Hu and Ji (2011) use topic, title, aspect and category words as their query terms. Li et al. (2011b) also introduce a graph-based sentence ranking by defining the category words and then using them for extracting information-rich sentences that will form the final summary.

Du et al. (2011) propose an aspect-based ranking model Decayed DivRank (DDRank) which aims at selecting top-ranked sentences by topic relevance that present topic-related, important, diverse and novel information contained in sets of documents. Steinberger et al. (2011) introduce an approach which aims to identify per-category aspects with the help of an event extraction system, to identify terms that are semantically related to the aspects with a system that automatically learns semantic classes. Steinberger et al. (2011)'s approach is not just based on extraction of the most salient sentences, but also on generation that aims to compress and reconstruct already extracted sentences in order to create a sequence of highly scored salient terms. The motivation for introducing compression into extractive multi-document summarization is that automatically generated summaries tend to contain longer sentences and the summary, therefore, tends to have on average fewer sentences than the summaries created by humans, which leads to less space in the summary. Compression allows to save that space and include more salient sentences and by that cover more content from the source documents.

Makino et al. (2011) also present a summarization model that is based on aspect coverage by using a maximum entropy classifier that predicts if each sentence contains information for the pre-defined aspects. This model calculates the scores that indicate aspect coverage, and the minimum of the aspects scores is then maximized for the summary to cover all the aspects. Therefore, this model is based on min-max problem and aims to provide a balanced coverage of aspects. Zhang et al. (2011b) introduce a summarization system which is also mainly concerned with the aspect coverage of summaries where aspect is recognized on the sentence level with the help of the aspect-bearing sentence recognition.

Cluster based approach: He et al. (2011) propose the category oriented extractive content selection approach which combines the evolutionary manifold ranking algorithm with spectral clustering based on eigenvector selection for computing the ranking score for each sentence, therefore, the summary is formed by selecting the sentences with the highest scores. The cluster based approach aims to group similar sentences to clusters that a sentence should belong to. Sentences that are highly related to a certain topic are grouped into one cluster. The top-ranked sentences in every cluster are selected to form the final summary. Zhang et al. (2011a) propose their PRIS summarization approach which is also based on sentence selection, where a topic-based sentence clustering algorithm and sentence ranking algorithm are employed to form the final summary by selecting the top ranked sentences in each cluster. Kennedy et al. (2011) introduce their uOttawa summarization system that is based on two modules: a clustering system which groups sentences based on topic and a sentence ranker that selects the sentence from each cluster closest to the query.

Kennedy et al. (2011) also propose to incorporate words describing emotions into the sentence ranking module, where emotional words are used as query expansion terms. The motivation to identify emotional categories within news articles is the hypothesis that the identified emotional words will help to select the sentences that will be more useful for being included into the summary. The summaries that contain emotional words are also believed to contribute to the readability of these summaries.

Das and Srihari (2011) combine corpus level (global) tag-topic models and target document set level local models in their extractive multi-document summarization approach. Chali et al. (2011) employ a query-focused extractive summarization approach based on a Markov chain model that follows a random walk paradigm for generating the most important sentences to produce the summary. Bakawid and Oussalah (2011) also employ a query-based extractive multi-document summarizer based on a Sentences Simplification Module (SSM) which aims to shorten the length of sentences via splitting or compression. Each sentence is then assigned a score according to its importance based on a set of features (overlap with the topic, concepts dominance and sentence position), then ranked based on their scores and the top n sentences are selected to form the summary.

The main problem of approaches based on content selection is that the selected sentences often turn out to be disconnected, unrelated to each other and therefore, are not able to form a coherent summary. Irrespective of how successfully the re-ordering algorithm works, these sentences are still rarely organized into an intelligible, readable summary.

2.2.3 Coherence Model for Multi-Document Summarization

While a great number of approaches used for multi-document summarization do not consider coherence in produced summaries as they focus mainly on salience and coverage, G-Flow (Christensen et al., 2013) introduces a novel system for extractive multi-document summarization that combines both selection and ordering tasks for producing coherent and salient summaries.

G-Flow's model is based on an approximate discourse graph where each vertex is a sentence from the input documents, and each edge represents a discourse relationship between sentences based on a number of indicators, such as deverbal nouns, lexical chains, discourse markers, the redundancy of information and coreference mentions.

G-Flow's summarization algorithm searches through the space of preliminarily ordered summaries and gives a score to each candidate summary with regard to coherence, salience and redundancy. Summary with the maximum value for the joint objective function that balances coherence, salience and redundancy and also takes into account the maximum summary length is picked by the G-Flow summarization system.

The G-Flow system's output summaries have been compared against four different state-of-the-art multi-document summarization systems, and the results of this comparison reveal that the joint model proposed by G-Flow outperforms all systems that are based on a pipeline of standard sentence selection and sentence reordering approaches. G-Flow summaries have also been evaluated against several quality dimensions used in DUC'04 like coherence of a summary, useless or repetitive text, handling entity mentions in text and other, and the quality of the G-Flow summaries have been significantly better than that of other systems. G-Flow's performance has also been rated as nearly the same as that of the human summaries (Christensen et al., 2013).

Chapter 3

Summarization Evaluation: Related Work

There has been a long history of research in the field of text summarization and its evaluation. Evaluation of automatic text summarization attempts to define how adequate a summary is in relation to its original text.

This section introduces the evaluation of various aspects of linguistic quality, which provides a strong background to understanding what makes a summary readable. This information, in its turn, contributes to the manual analysis and identification of various linguistic violations that have been detected in automatically generated summaries and also to further development of the annotation scheme.

3.1 Evaluation Types

Summary evaluation methods are typically divided into two types: intrinsic and extrinsic (Hahn and Mani, 2000). The intrinsic type of evaluation is based on evaluating the quality of the summary by directly analyzing this summary, while extrinsic evaluation is based on evaluating the quality of a summary by judging how this summary completes some other task, for example a reading comprehension task where the evaluation is based on whether the answer is equivalent to the answer taken from the source text (Hovy, 2005; Mani, 2001).

Most evaluations of text summarization systems are based on intrinsic evaluation methods (Hovy, 2005). Intrinsic evaluation methods are often fulfilled by comparing summaries to some gold standard or relying on human judgments of the goodness of a summary. Intrinsic evaluation mainly assesses the coherence and informativeness of summaries. Coherence is responsible for a summary's well-organized structure, and informativeness aims at assessing the summary's information content by measuring content overlap between an ideal summary and the output of a summarization system (Mani, 2001).

3.2 Human Evaluation

Evaluation of summaries traditionally includes human judgments of the quality of final summaries. Human evaluation is known to be a complex cognitive process and, inevitably tends to cause certain difficulties. Indeed, the task of manually evaluating summaries and systems which produce these summaries is not straightforward. The main difficulty is that there is no ideal summary for a given document or a set of documents. An ideal summary is hard to create and it is rarely unique. Since all humans perceive information differently, there will be many different ways to describe one and the same event. Different assessors will generate different summaries of the same source that they consider acceptable (Hahn and Mani, 2000). Additionally, no single summary generated by a human can be considered as an ideal one because there will always be disagreement between human assessors, both for generating and for evaluating summaries (Das and Martins, 2007; Mani, 2001). This demonstrates the instability of manual evaluation and confirms that using a single model as a reference summary is not appropriate (Lin and Hovy, 2002). Besides that, manual evaluation of summaries is very expensive requiring a lot of time and great human effort.

3.3 Automatic Content Evaluation

In order to overcome the instability of human evaluation and to reduce costs, considerable attention has been devoted to automatic summary evaluation. As automatic methods for evaluating the quality of texts generated by machines can be applied at development time they have gained popularity. Due to their fast applicability and reliability some such methods are still followed today as foundations for evaluating text summarization.

3.3.1 ROUGE

One of the most prominent works which made a valuable contribution to the field of evaluation of text summarization is Lin (2004) research where he suggested an automatic evaluation method for evaluating the quality of content selection called ROUGE. ROUGE is widely used as it greatly reduces the complexity of evaluations and therefore, is cheap and fast to implement (Louis and Nenkova, 2009).

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. ROUGE has become a standard for automated evaluation of summaries that represents a set of metrics for automatic evaluation of the quality of a summary. These metrics measure the similarity between summaries by comparing them to a model or ideal summary manually written by humans (Lin, 2004). Such measures count the number of overlapping units such as n-grams, word sequences, and word pairs between the automatically generated summary and the summary created by humans.

The results of the comparison made by ROUGE are evaluated using precision, recall and F-measure. Precision and recall are considered to be standard measures which reflect how many of the system's extracted sentences were good and how many good sentences the system missed (Hovy, 2005). Precision and recall measures are combined in a so-called F-measure which simply calculates a weighted mean of precision and recall.

3.3.2 Pyramid method

In recent years, the Pyramid evaluation method has been introduced to improve the content evaluation of summaries. The Pyramid method is a reliable approach providing stable evaluation scores that lead to improved content evaluation. This method is introduced to avoid the problems related to comparing automatic summaries with a single gold standard and is based on semantic analysis of multiple human gold standards. This evaluation process makes use of multiple model summaries although it is still very hard to estimate how many model summaries are needed to achieve reliable automatic summary evaluation.

Nenkova and Passonneau (2004) introduce Summary Content Units (SCU) and mark information that semantically matches as expressing the same SCU over distributions of human summaries. Pyramid evaluation measures content selection with the following procedure: first, four model summaries for the document set are created, then the assessor extracts Summary Content Units (SCUs) from these multiple summaries and sorts the content units into various aspect bins (one bin per aspect of a given category). Each SCU has a weight equal to the number of model summaries in which assessors found a content unit. Once all SCUs have been collected from the model summaries, the assessor determines which of these SCUs can be found in each of the peer summaries that are to be evaluated. Each SCU contained in the peer summary is counted only once. The final Pyramid score for a peer summary is the ratio between the weights of SCUs contained in a summary, and the SCU weights of a perfectly informative summary with the same number of SCUs (Lin and Hovy, 2002; Nenkova and Passonneau, 2004; Nenkova et al., 2007; Nenkova and McKeown, 2011).

3.4 Linguistic Quality Evaluation

This section describes different approaches to automatic evaluation of linguistic quality and also introduces a number of aspects of linguistic quality that are important for summaries. Having such a great variety of aspects of linguistic quality it is difficult to state for sure which aspects of linguistic quality are more important for summary's readability. The previous work described in this section gives a grounding for the manual analysis of various violations of linguistic quality with the purpose of further manual annotation of summaries.

Both ROUGE and the Pyramid method described in the previous section are measures for evaluating the content of a summary (informativeness), but not readability. Readability is an important factor in evaluating the overall quality of summaries as it helps to present the useful content in a coherent and structured way, making final summaries easier to read and understand.

Readability is frequently described as a combination of five aspects of linguistic quality: grammaticality, non-redundancy, referential clarity, focus, and structure/coherence.[3] Measuring these aspects of linguistic quality of summaries does not involve comparison with a model summary, but only concerns the automatically produced output summaries. The linguistic quality definitions are used by humans to assess the readability and fluency of output summaries. It is important to point out that humans do not consider the relationship between the output summary and the input documents, but aim to evaluate the final summary as a document in its own right.

The exact definitions of each aspect given by the Document Understanding Conference (DUC) are reproduced below:

- *Grammaticality*: The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.

- *Non-redundancy*: There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., "Bill Clinton") when a pronoun ("he") would suffice.

- *Referential clarity*: It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.

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

- *Focus*: The summary should have a focus; sentences should only contain information that is related to the rest of the summary.
- *Structure and Coherence*: The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

3.4.1 Automatic Evaluation of Local Coherence

A number of studies have been pursued in the field of evaluation of different aspects of linguistic quality of summaries. Most of these studies focus on a single linguistic factor that contributes to readability. Among various aspects of readability, structure/coherence remains one of the most challenging factors in summary evaluation (Pitler et al., 2010).

Barzilay and Lapata (2005) and Lapata and Barzilay (2005) address the problem of automatic evaluation of *local coherence* and its correlation with human judgments. Local coherence focuses on capturing relatedness between sentences that form a globally coherent text. An entity-based representation of discourse used by Barzilay and Lapata (2005) for coherence assessment measures coherence by reflecting important patterns of sentence-to-sentence transitions. This turns out to be a useful metric for evaluating summaries. Lapata and Barzilay (2005) implement two classes of coherence models - one based on the *syntactic* aspect and another one on the *semantic* aspect of text coherence. The syntactic coherence model tracks all mentions of the same entity in different syntactic positions which are spread across adjacent sentences. The semantic model in its turn measures the semantic relatedness between adjacent sentences. These two aspects for coherence evaluation are complementary; therefore, their fusion demonstrates a more significant agreement with human ratings than each individual approach (Lapata and Barzilay, 2005).

Coreference resolution as the process of finding all references to the same entity in a text is also important for evaluation of local coherence, which in its turn contributes to text readability. Nenkova and McKeown (2003) analyze the syntactic forms of noun phrases that represent references to named entities and emphasize the difference between syntactic forms of entities that are introduced in a text for the first time - first mentions - and those of subsequent mentions. Elsner and Charniak (2008) ignore the syntactic form of references and exploit coreference resolution for coherence modeling. The coreference-inspired models used by Elsner and Charniak (2008) achieve a significant improvement over the entity-grid proposed by Lapata and Barzilay (2005).

Grammaticality does not seem to be much of an issue for extractive summaries as sentences are picked from the original document with little or no modification. Nevertheless, Vadlapudi and Katragadda (2010) explore how the evaluation of *grammaticality* and *struc-*

ture/coherence can be automated. Summary grammaticality is measured with the help of a POS n-gram model that estimates the probability of a sentence being grammatically acceptable. The assumption for this estimation is that sentences constructed by applying frequently-used grammar rules would have higher probability and, therefore, form well-accepted sentences. Coherence makes text semantically meaningful, and in order to capture this lexical cohesion in summaries, Vadlapudi and Katragadda (2010) make use of lexical chains that represent topics that are being discussed throughout a text.

3.4.2 Automatic Evaluation of Sentence Fluency

Nenkova et al. (2010) consider *sentence fluency* to be an important factor of the overall linguistic quality of text and therefore explore its predictors. Sentence fluency depends on many criteria, one of them being vocabulary use. The vocabulary used in a sentence or text largely influences its readability. If a text contains rare difficult words or special terms that are not used daily by the reader, it will be perceived as less readable. In general, more commonly used words are easier to read and understand. Nenkova et al. (2010) explore feature classes which might turn out to be predictive of good fluency. They introduce a great number of such features, but I will mention just two of them. One of the features that could be strongly associated with fluency is sentence length. The motivation for this is that shorter sentences are easier to process and thus shorter sentences are more likely to be perceived as fluent. The other feature type is the parse tree depth, which measures how complex a sentence is and how many noun phrases, verb phrases and prepositional phrases it contains. The motivation for this is that longer sentences tend to be syntactically more complex and therefore may slow processing and lead to lower perceived fluency of the sentence.

3.4.3 Automatic Evaluation of Multiple Aspects of Linguistic Quality

Pitler and Nenkova (2008) unify lexical, syntactic and discourse features to produce a highly predictive model of human readers' judgments of text readability and see how this combination affects the perceived text quality. They also analyze such readability factors as vocabulary, syntax, cohesion, entity coherence and discourse, where discourse relations play a special role in evaluating the perceived quality of text.

Pitler et al. (2010) present numerous sets of linguistic features for automatic evaluation of linguistic quality of summaries and attempt to identify the best feature classes for various aspects of text quality. Such feature classes as word choice and word coherence, the reference form of named entities, entity coherence, cohesive devices, continuity and sentence fluency have been examined and further evaluated for predictive power of the accuracy of certain aspects of linguistic quality.

Chapter 4

Datasets

This section introduces two datasets, TAC and G-Flow, which will provide the basis for producing an annotated corpus of linguistic quality violations that occur in automatically generated summaries.

4.1 TAC Dataset

The data used for this project comes from the TAC 2011 Guided Summarization Task which is a multi-document summarization task where various methods and results are compared on shared test data sets.

The documents used for this summarization task are taken from the newswire section of the TAC 2010 KBP Source Data (LDC Catalog Number: LDC2010E12). The collection of documents covers the years 2007-2008 and consists of news articles taken from the New York Times, the Associated Press, and the Xinhua News Agency newswires.

The aim of the TAC 2011 Guided Summarization Task was to write a 100-word summary of a set of 10 news articles for a certain topic. There are 44 topics in the dataset. Each topic may fall into one of five predefined categories:[4]

1. Accidents and Natural Disasters
2. Attacks
3. Health and Safety
4. Endangered Resources

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

5. Investigations and Trials

The participants of the task were given a list of important aspects for each category, and a summary was supposed to cover all these aspects, obviously only if the information was found in the documents.

The dataset includes the multi-document summaries that were produced by humans, so-called model summaries - D1101-A.models and the summaries generated automatically by summarization systems ("peers"), so-called D1101-A.peers. For each collection of documents of a certain topic there are 4 model summaries manually created by humans and 50 summaries automatically generated by various summarization systems. Some summarization systems produce abstractive types of summaries, which will not be included into this project.

The collections ending with "-A" are the ones for the "Guided Summarization Task", which is essentially an aspect-based multi-document summarization, where an aspect is some sort of a query used in summarization. The collections ending with "-B" are the ones for the "Update Summarization Task", which requires the generation of an update summary assuming that the reader has already read some articles related to this topic before. Both "-A" and "-B" sets are of the same topic but "-B" set contains documents which have been published later than those in "-A". For this project we consider only documents for the Guided Summarization Task and only those summaries that have been automatically generated by summarization systems, which we will refer to as "document collection A.peers".

4.2 TAC Data Manual Evaluation

As required by the TAC challenge all automatically generated summaries get truncated to 100 words, then both human-written summaries and automatically generated summaries are manually evaluated by humans.

TAC intends not only to capture the performance of content selection but also to address the problems of readability or other text qualities of summaries. The quality of each summary in this dataset (those created by humans and those generated automatically by summarization tools) has been evaluated for (Nenkova, 2005):

1. Content
2. Readability/Fluency
3. Overall responsiveness

Readability measures linguistic quality and has been evaluated for its five factors: grammaticality, non-redundancy, referential clarity, focus, and structure/coherence.

An Overall Responsiveness score is based on both content and readability/fluency and measures not only how well a summary provides information relevant to the topic defined by the user, but also the overall linguistic quality of the final summary.

Readability and Overall Responsiveness are each rated by human assessors on a 1 to 5-point scale, where 5 is the best score. These ratings demonstrate how good a summary is to an assessor in terms of providing relevant information and producing a fluent summary.

Readability and Overall Responsiveness assessments are done without reference to any model summary and are fully based on human judgments.

4.3 G-Flow Dataset

The G-Flow dataset used for this project comes from the Document Undestanding Conference (DUC)[5] for 2004 with the setting of Task 2, which is a multi-document summarization task on English news articles. The documents for Task 2 consist of news articles taken from Associated Press Newswire, New York Times Newswire, and Xinhua News Agency.

In this summarization task, 50 clusters of related documents are provided, each of which consists of 10 documents. Each document cluster also includes four gold standard summaries.

The G-Flow dataset includes 50 short multi-document summaries generated for each document cluster. As required in the DUC'04 evaluation, a single summary should not be longer than 665 bytes including spaces and punctuation (Christensen et al., 2013).

[5] http://duc.nist.gov/

Chapter 5

Linguistic Quality Violations: Analysis and Annotation

In this chapter we discuss the analysis of the summaries and present the description of the annotation scheme for creating linguistically annotated corpora for various violations of linguistic quality in multi-document summarization. We also provide an overview of the set of annotation tags used for this annotation task followed by individual descriptions of each tag, various attribute values that are available for these tags and examples from the TAC dataset. The annotation tool used for the manual annotation, the annotation process and the measurement of the inter-annotator agreement are also discussed in this chapter.

5.1 Annotation Scheme

The process of corpus annotation includes creating annotation spans, which carry the annotation information. The attributes assigned to the defined annotations describe different types of violations of linguistic quality. In other words, a certain type of annotation is assigned to a span. Four annotation types out of five contain attributes that specify different violation types.

Each type of violation occurring in summaries is marked for annotation. For example, the sentence below, taken from a summary contains a violation *CHD* that gets assigned with a span with the starting character offset = "483" and ending character offset = "486". This span gets assigned the annotation of *entity mention*, with an attribute of *entity mention violation* of value *acronym without explanation*.

> "This is the first study to measure the changing number of patients in a North American population during a period of major progress in the management of <u>CHD</u>," said

> Ariane J. Marelli, M.D., lead author of the studyat McGill University in Montreal, Quebec, Canada.

As shown in Figure 5.1 at the core of the annotation scheme is a set of 5 classes of annotations:

1. entity mention
2. relation between entities
3. clause
4. relation between clauses
5. misleading discourse connective

All annotation types have attributes. The annotation types *entity mention* and *clause* have two types of attributes, one that allows selecting a value from a predefined list and the other - *comment* - that allows for the value to be entered as free text. The annotations *relation between entities* and *relation between clauses* have only one attribute with values that can be selected from the predefined list. The annotation *misleading discourse connective* has also one attribute - *comment* - that allows to enter a value as text. All attribute values represent different violations of linguistic quality and the total number of violation types is 16 among which 6 types represent relations between annotations.

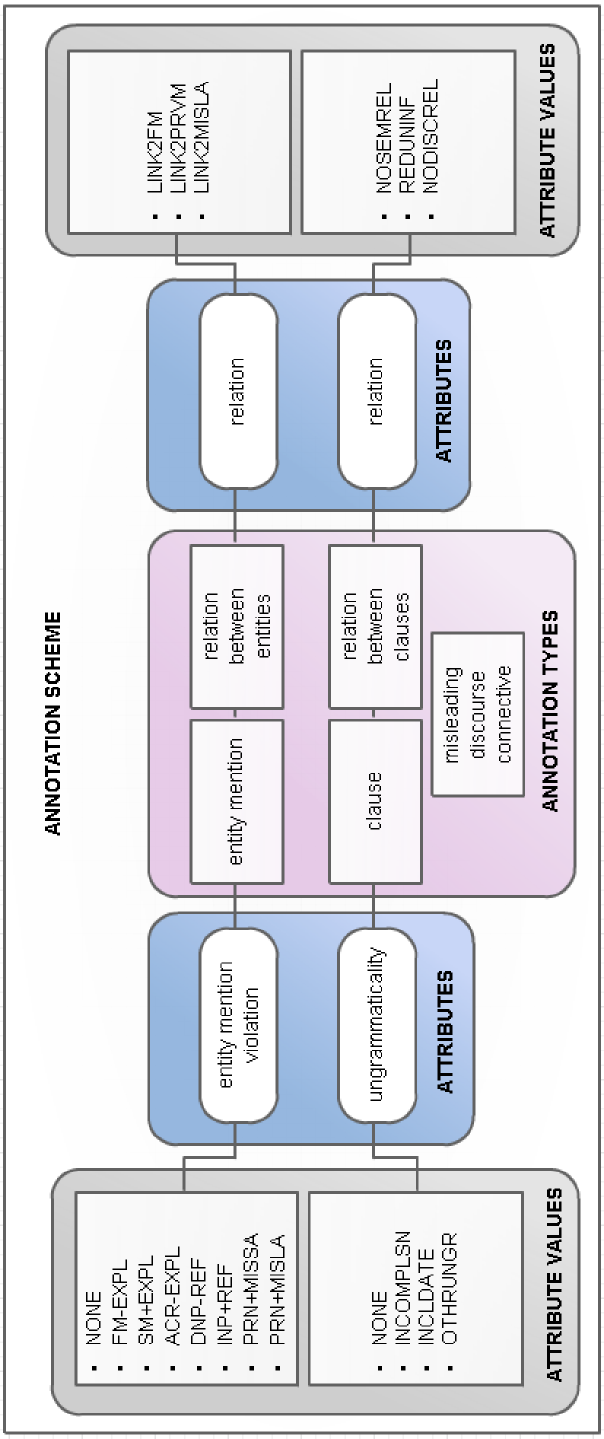

Figure 5.1. Annotation scheme for factors of linguistic quality, with attributes and their values

Table 5.1 provides the abbreviations of the violation types for each annotation tag, which will be used in the following chapters.

Violations for each annotation tag	Abbreviation
entity mention	
first mention without explanation	FM-EXPL
subsequent mention with explanation	SM+EXPL
acronym without explanation	ACR-EXPL
definite np without reference to previous mention	DNP-REF
indefinite np with reference to previous mention	INP+REF
pronoun with missing antecedent	PRN+MISSA
pronoun with misleading antecedent	PRN+MISLA
relation between entities	
link to first mention	LINK2FM
link to previous mention	LINK2PRVM
link to misleading antecedent	LINK2MISLA
clause	
incomplete sentence	INCOMPLSN
inclusion of datelines	INCLDATE
other ungrammatical form	OTHRUNGR
relation between clauses	
no semantic relatedness	NOSEMREL
redundant information	REDUNINF
no discourse relation	NODISREL
misleading discourse connective	
misleading discourse connective	MISLDISCON

Table 5.1. List of abbreviations for the violation types for all annotation tags

5.1.1 Annotation of Entity Mentions

An entity is an object or a set of objects in the world that represent individuals, organizations and locations. An entity mention is a reference to an entity. Entities may be referenced in a text by their names, indicated either by a common noun or a noun phrase, or represented in a text as a pronoun.

The annotation *entity mention* is used to annotate entity mentions in summaries. Only entity mentions that have certain violations or entities that do not have any violations but stand in a relation with other entities which have been violated will be used for annotation. Once the entity mention has been annotated by selecting the *entity mention* annotation type, now it is possible to set values for the attributes of the annotation. The *entity mention violation* attribute may take one of the values defined below:

Attribute Value: none The default attribute value none is used when an entity has no violation related to entity mentions, but will be further used for setting a relation between two entity tags.

In example (1a) *the Adam Air Boeing* is an entity mention that has no violations. Even if this entity has no entity mention violation, it will still be selected for annotation and will be assigned a value *none* as it will be used for setting a relationship with another violated entity *an Adam Air plane* in (1b).

(1) (a) The Adam Air Boeing 737-400 crashed afternoon, but search and rescue teams only discovered the wreckage early next morning. (b) Three Americans were among the 102 passengers and crew on board an Adam Air plane which crashed into a remote mountainous region of Indonesia.

Attribute Value: first mention without explanation (FM-EXPL) The attribute value *first mention without explanation* is used to mark entity mentions which are introduced in the text for the first time and lack brief informative descriptions. This type of annotation will include only uncommon named entities representing individuals, organizations and locations. Named entities which are well-known to an average reader (for example, President Obama) and that miss a brief explanation in a text will not be selected for annotation.

Whenever a non-generally-known named entity is mentioned for the first time in text and a brief description of this first mention is not provided, the text tends to get less readable making the reader have problems with understanding the person or location the summary is going to be about. The provided description of the unfamiliar entity, therefore, should improve the perceived readability of the text.

In example (2) the named entity *Roberts* is introduced in the summary for the first time, but the description about who this person might be is not provided.

(2) Roberts killed himself in the one-room remote Amish schoolhouse before police could get to him, State Police Commissioner Jeffrey Miller told a press conference.

Attribute Value: subsequent mention with explanation (SM+EXPL) The attribute value *subsequent mention with explanation* is used to mark subsequent mentions of entities which come with a brief informative description in a text while their first mentions do not have an informative description.

In example (3b) the named entity *Tony Taylor* is a subsequent mention of the person. This subsequent mention contains a brief description of who this person is, by providing information on his age and where he is coming from. Since this person has already been introduced into discourse by (3a), *Tony Taylor* in (3b) is an entity mention violation, where a subsequent mention has too much information, while the first mention has no explanation.

(3) (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 9 a.m.

Attribute Value: acronym without explanation (ACR-EXPL) The attribute value *acronym without explanation* is used to mark acronyms which lack the information describing what these acronyms stand for. Only acronyms which come with a description in an original document but lack this description in a summary should be selected for annotation. Well-known acronyms such as WWF, US, HIV and CIA will be excluded from the annotation process due to their familiarity to an average reader.

Example (4) contains an acronym *CITES* which has been explained in the original document but misses the explanation in the summary. CITES stand for the Convention on International Trade in Endangered Species and is unlikely to be known to an average reader.

(4) "Red coral is the most valuable and widely traded out of all the coral species, and CITES protection will help ensure the future of the species and the red coral industry," he said.

Attribute Value: definite noun phrase without reference to previous mention (DNP-REF) Definite noun phrases indicate that they should be known to a reader from the context. The attribute value *definite noun phrase without reference to previous mention* is used to mark definite noun phrases. However, these noun phrases lack references to entities which should have been previously mentioned in text.

In example (5) it is unclear which girls the definite noun phrase *the girls* is referring to as these girls have not been mentioned in the text before.

(5) Roberts killed himself as police stormed the building. In Lancaster County, there have been prayer services for the Amish school shooting victims at area churches, but the traditional funerals for the girls were closed.

Attribute Value: indefinite noun phrase with reference to previous mention (INP+REF) The attribute value *indefinite noun phrase with reference to previous mention* is used to mark indefinite noun phrases that refer to the same entity as a previous entity mention in the summary. The use of indefinite noun phrases in text when this entity has already been introduced into discourse earlier is a violation from the linguistic point of view.

In example (6b) the use of the indefinite noun phrase *an Adam Air plane* is marked as a violation as it refers to the same entity which has been previously mentioned in sentence (6a).

(6) (a) The Adam Air Boeing 737-400 crashed afternoon, but search and rescue teams only discovered the wreckage early next morning. (b) Three Americans were among the 102 passengers and crew on board an Adam Air plane which crashed into a remote mountainous region of Indonesia.

Attribute Value: pronoun with missing antecedent (PRN+MISSA) The attribute value *pronoun with missing antecedent* is used to mark pronouns in text which miss antecedents. Both personal and possessive pronouns are annotated.

Pronoun coreference is an important aspect of coherence. A coreference chain often includes pronouns and common nouns that refer to the same person. It is important to place pronouns close to appropriate referents with the correct number and gender. If a pronoun that refers to the same person is too far away from its referent, it becomes hard to interpret it. Complete absence of a referent creates confusion, and having too many referents creates ambiguity. Missing antecedents in a summary make it impossible to understand who these pronouns are referring to.

In example (7b) the pronoun *he* does not have an antecedent in the preceding sentence (a). Therefore, it becomes unclear who this person is referring to in the summary.

(7) (a) 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. (b) He is charged with 191 counts of murder and 1,755 of attempted murder in the Madrid, Spain, bombings on March 11, 2004, though he has been secretly recorded saying that he had not been with the men who carried them out.

Attribute Value: pronoun with misleading antecedent (PRN+MISLA) The *pronoun with misleading antecedent* attribute value is used to mark pronouns in text which have an antecedent which is misleading. An antecedent is misleading if it is not what the annotated pronoun should be referring to according to the original text. Both personal and possessive pronouns are annotated.

In example (8) in sentence (c) it is unclear who the pronoun *they* is referring to. One could think that *they* in sentence (c) refers to the *police* in sentence (b). In fact, *they* refers to *young relatives* that Roberts claimed he had molested decades ago in (a). Therefore, the antecedent *police* in (b) is a misleading antecedent as it is not the one that the pronoun *they* should be referring to.

(8) (a) Roberts had revealed to his wife in a note left behind the day of the attacks and in a cell phone call from inside the school that he was tormented by memories of molesting two young relatives 20 years ago and dreamed of molesting again. (b) It was unclear if the shooter was among the six, but state police had said earlier that he had been killed. (c) They were absolutely sure they had no contact with Roberts.

5.1.2 Annotation of Relations between Entities

The class level tag *relation between entities* is used to identify relations between two entity mentions. This type of annotation contains only one attribute called *relation* which may take one of the values described below.

Attribute Value: link to first mention (LINK2FM) The attribute value *link to first mention* is used to mark the relation between a subsequent mention of an entity that appears in the text with an informative description and the first mention of the same entity.

In example (9) sentence (b) contains the subsequent mention of a person that comes with more information about this person, while sentence (a) contains the first mention of the same person but lacks the informative description.

(9) (a) Roberts killed himself in the one-room remote Amish schoolhouse before police could get to him, State Police Commissioner Jeffrey Miller told a press conference. (b) On Monday morning, Charles Carl Roberts IV entered the West Nickel Mines Amish School in Lancaster County and shot 10 girls, killing five. It was the first in a series of funerals Thursday for victims of the West Nickel Mines Amish School shooting.

In example (10) sentence (b) contains a subsequent mention of a named entity which comes with a brief informative description, while the first mention of the same entity in sentence (a) has no such description.

(10) (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 9 a.m.

Attribute Value: link to previous mention (LINK2PRVM) The attribute value *link to previous mention* is used to mark the relation between an indefinite noun phrase representing an entity and a previous mention of the same entity.

In example (11) indefinite noun phrases a *gunman* and *a schoolhouse* in sentence (b) clearly reference the same entity as the entity mention in the preceding sentence (a) and therefore, the use of indefinite noun phrases is incorrect.

(11) (a) On Monday morning, Charles Carl Roberts, entered the West Nickel Mines, the Amish School in the County and shot just 2 for 10 in the girls at the time of the killing. b) People were dead after a gunman opened fire at a one-room Amish schoolhouse on Monday, Pennsylvania's Lancaster County County coroner said.

In example (12) the indefinite noun phrase *an Adam Air plane* in sentence (b) clearly references the previous entity mention *the Adam Air Boeing* in the preceding sentence (a).

(12) (a) The Adam Air Boeing 737-400 crashed afternoon, but search and rescue teams only discovered the wreckage early next morning. (b) Three Americans were among the 102 passengers and crew on board an Adam Air plane which crashed into a remote mountainous region of Indonesia.

Attribute Value: link to misleading antecedent(LINK2MISLA) The attribute value *link to misleading antecedent* is used to mark the relation between a pronoun with a misleading antecedent and an entity which is selected as the misleading antecedent.

In example (13) in sentence (c) the pronoun *they* has a misleading antecedent the *police* in sentence (b). The correct antecedent for the pronoun *they* is *relatives* in sentence (a).

(13) (a) Roberts had revealed to his wife in a note left behind the day of the attacks and in a cell phone call from inside the school that he was tormented by memories of molesting two young relatives 20 years ago and dreamed of molesting again. (b) It was unclear if the shooter was among the six, but state police had said earlier that he had been killed. (c) They were absolutely sure they had no contact with Roberts.

In example (14) in sentence (b) pronoun *it* refers to the misleading antecedent *the Las Vegas Springs Preserve.*

(14) (a) One way water officials are trying to do that is through the Las Vegas Springs Preserve.(b) It would be in effect through 2026 and could be revised during that time.

5.1.3 Annotation of Clauses

The annotation type *clause* is used when clauses contain violations of coherence. We annotate both clauses containing violations and clauses that stand in some sort of problematic relation with another clause. The *clause* annotation has two types of attributes, one is *comment* for entering values as free text and the other is *ungrammaticality* which can take one of the following values:

Attribute Value: none The default attribute value *none* is used when there is no violation within a clause; instead this violation is represented through the relation with another clause.

In example (15) both adjacent sentences (a) and (b) are sentences that appear in the same summary and they do not have any violations within them, but they are still not semantically related to one another.

(15) (a) Lewis had no major endorsements at the time, only a shoe deal and some local contracts. (b) Tony Taylor, 34, of Hampton, Va., has a plea-agreement hearing scheduled for 9 a.m.

Attribute Value: incomplete sentence (INCOMPLSN) The attribute value *incomplete sentence* is used to mark incomplete sentences. The main reason for summaries to have incomplete sentences is the restriction of summary length. Some summaries get truncated to 100 words, which sometimes makes final sentences of long summaries suddenly break without completion. Incomplete sentences may also occur in summaries due to some system-internal formatting.

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

Attribute Value: inclusion of datelines (INCLDATE) The attribute value *inclusion of datelines* is used to mark phrases which contain datelines typical of news articles that usually provide information about the location, date and time of the described event.

This information is usually included in the very first sentence of a news article. The presence of such datelines in a summary tends to disrupt the fluency of the summary.

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

Attribute Value: other ungrammatical form (OTHRUNGR) The attribute value *other ungrammatical form* is used to mark phrases/clauses which contain ungrammatical forms such as missing spaces between words or other types of grammatical errors. A summary in example (18) contains sentence (a) which has a comma at the end of the sentence instead of a full stop, sentence (c) which does not have any punctuation mark at the end of the sentence, and sentence (d) which does not have a space between two words.

(18) (a) A gunman shot girls in the head execution style at an Amish school in Pennsylvania state on Monday, killing four wounding at least six others, (b) Two relatives of the man who attacked Amish school girls said Monday they were not molested by him 20 years ago, as he had claimed in a phone call to his wife during the siege. (c) Police say shooter at Amish school told wife he molested years ago, dreamed of doing it again (d) On Monday morning, Charles Carl Robertsentered the West Nickel Mines Amish School in Lancaster County shot 10 girls, killing five.

5.1.4 Annotation of Relations between Clauses

The annotation *relation between clauses* is used for indicating a relation between two clauses. This annotation contains only one attribute called *relation.* The *relation* attribute may take one of the values described below.

Attribute Value: no semantic relatedness (NOSEMREL) The attribute value *no semantic relatedness* marks the relation between two adjacent clauses that are semantically unrelated to each other.

In example (19) sentences (a) and (b), and sentences (c) and (d) demonstrate that there is no semantic relatedness between two adjacent sentences.

(19) (a) Charles Carl Roberts IV, 32, held a steady job working nights driving a truck that collected milk from area dairy farms. (b) Police could offer no explanation for the killings. (c) The shootings occurred about 10:45 a.m. (d) Firewood and children's toys, including two play guns, were on the porch.

Attribute Value: redundant information (REDUNINF) The attribute value *redundant information* is used to mark phrases or sentences which express the same information.

One of the problems with multi-document summarization of news, where the input is a set of articles on the same topic, is that sentences selected to form a final summary often express the same meaning. Summaries created by humans usually avoid any redundant information, while automatic summarization systems generate summaries which contain one or more alternate sentences that express the same meaning and as a result become less readable.

In example (20) both highlighted phrases convey the same information and therefore, stand in a redundancy relation.

(20) 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.

Attribute Value: no discourse relation (NODISREL) The attribute value *no discourse relation* is used to mark two adjacent sentences that do not stand in a discourse relation with each other only if there is a misleading discourse connective in one of the sentences. This type of relation shows that two segments of discourse are logically not connected to one another.

Discourse relations are often explicitly signaled, which means they are easily identifiable in discourse with the help of discourse connectives which are present overtly in the text. Discourse relations are associated with two adjacent sentences. The adjacent sentences become arguments of a discourse relation, where one of the arguments is syntactically related to an explicit discourse connective. Explicit discourse connectives in a text can be represented as subordinating conjunctions (*because*, *since*, *though*), coordinating conjunctions (*and*, *or*) or adverbials (*however*, *otherwise*, *then*). These discourse connectives help to infer the discourse relations between adjacent sentences of the text, and the lack of proper context for the use of discourse connectives makes summaries incoherent.

In example (21) sentences (a) and (b) do not stand in a discourse relation that the connective invokes.

The explicit discourse connective *and* lacks the proper context for the use of this discourse connective. If these two sentences were logically connected to each other, the discourse connective would have helped to infer the discourse relations between these two adjacent sentences.

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

5.1.5 Annotation of Misleading Explicit Discourse Connectives

The annotation *misleading discourse connective* (MISLDISCON) is used for annotating misleading explicit discourse connectives. Only misleading discourse connectives that have been inappropriately used in a summary and show the absence of the discourse relation between adjacent sentences should be selected for annotation. Explicit discourse connectives in a text can be represented as subordinating conjunctions (for example, *because*, *since*, *though*), coordinating conjunctions (for example, *and*, *or*) or adverbials (for example, *however*, *otherwise*, t*hen*).

In example (22) sentence (a) is the first sentence in a summary that contains the discourse connective *then* and which lacks the proper context and therefore signals the absence of the discourse relation between two adjacent sentences (a) and (b).

(22) (a) Then in 1935, the Hoover Dam opened on the Colorado River. (b) Arizona water managers, called for the plant to be turned on.

5.2 Annotation Process

Annotation has been carried out on extractive summaries only, while 176 abstracts from the TAC dataset - 4 abstracts generated by 2 summarization systems with system IDs 38 and 49, 36 and 50 in each document cluster - have been excluded from the annotation process. Also, 88 extractive summaries - 2 summaries per each document cluster of one summarization system with system IDs 27 and 30 are missing from the dataset and therefore, have not been annotated. Otherwise, all the TAC data has been annotated.

There are 44 TAC document clusters that contain around 44 summaries related to one topic. The annotation process has been organized in a way that only 5-8 summaries per cluster are annotated at a time. This ensures annotation consistency and allows for the annotator not to get overloaded with information on the same topic.

5.3 Annotation Tool

Annotation is carried out using MAE (Multi-purpose Annotation Environment)[6], an annotation tool for manual linguistic annotation of texts.

Employing the MAE annotation tool for this project is suitable because it allows for arbitrary spans of annotation and also easily creates links between annotations (Stubbs, 2011). MAE defines the text spans by using character offsets where each annotation has the start-offset which indicates the first character of the annotated span in the text and the end-offset that specifies the first character after the annotated span. This manner of defining spans is beneficial for this annotation task as due to some summarization systems' internal formatting, some texts in the TAC dataset contain sentences with no spaces between words, hence token-based annotation is impossible.

For this annotation task, MAE has slightly been customized by adding the visualization of the relations between annotations by means of arcs between spans.

5.4 Inter-Annotator Agreement

In order to test the reliability of the annotation scheme and measure the level of agreement between annotators, the same set of 100 summaries have been annotated by 2 annotators, further referred to as annotator 1 and annotator 2. The subset has been picked at random and contains 95 TAC and 5 G-Flow summaries. The full list of the files selected for measuring the inter-annotator agreement can be found in Appendix A.

Annotators were provided with a detailed annotation manual that contained descriptions of each annotation tag and its possible attributes followed by examples. Prior to commencing the annotation process, the annotators tested both the application of the annotation scheme and the use of the MAE annotation tool. There was also a training phase, when annotators tried annotating 20 randomly selected documents (15 TAC and 5 G-Flow summaries).

Originally, to measure the inter-annotator agreement the same subset of 100 extractive summaries have been annotated by 3 annotators, but annotator 3 has been excluded from the analysis of the results due to the low agreement scores with the primary annotator (annotator 1). The counts of all types of violations and the scores of Precision, Recall and F-score for annotator 3 are in Appendix C.

The agreement between two annotators includes the cases when two annotators agree not only on the annotation tag and a violation type (attribute), but also on the beginning

[6] https://code.google.com/p/mae-annotation/

and the end of the spans selected for the annotation of a phenomenon. Two spans are considered to be fully matched if the start and end of one annotation match the start and end of another annotation. Figure 5.2 shows that the start and the end of annotations A and B are fully matching.

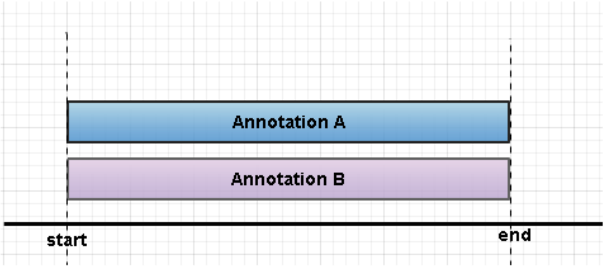

Figure 5.2. The case of full match between two annotations

Spans that are overlapping or partially match have also been considered as agreement between two annotators. Two spans are considered to match partially if:

(1) the start of one annotation is within the start and the end of another annotation. Figure 5.3 shows the cases when the start of annotation A is within the start and end of annotation B.

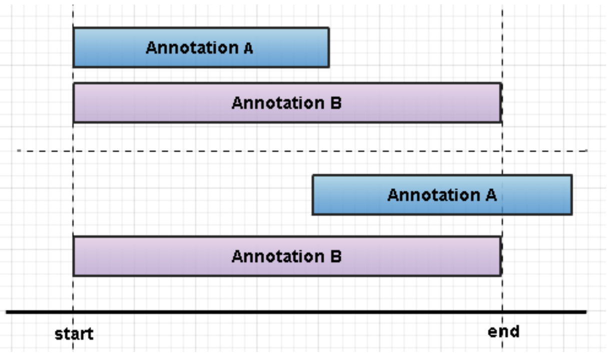

Figure 5.3. Possible cases of partial match between two annotations (1)

(2) the end of one annotation is within the start and the end of another annotation. Figure 5.4 demonstrates the cases when the end of annotation A is within the start and the end of annotation B.

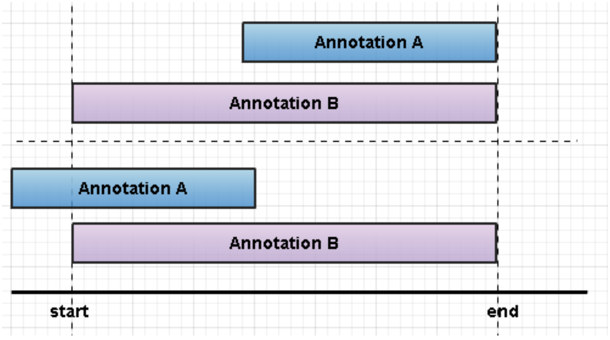

Figure 5.4. Possible cases of partial match between two annotations (2)

(3) the start and the end of one annotation is within the start and the end of another annotation. Figure 5.5 shows the possible cases when the start and the end of annotation A is within the start and the end of annotation B or vice versa, the start and the end of annotation B is within the start and the end of annotation A.

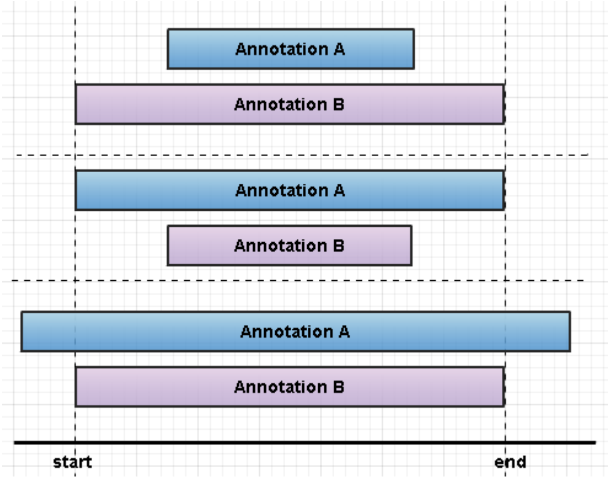

Figure 5.5. Possible cases of partial match between two annotations (3)

Relations between two annotations match when the relation type matches and also the annotation spans that are selected to be in this relation also fully or partially match. The disagreement between two annotators takes place when violation types (attribute values) mismatch.

The first annotator in each pair of annotators has been treated as gold standard or reference for measuring the extent to which the other annotator deviates from this reference. The standard evaluation metrics Precision, Recall and F-score have been calculated for each type of violation across all annotation tags and for each pair of annotators. The

F-score is the traditional F_1 score that combines Precision and Recall by producing the harmonic mean of the two:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5.1}$$

The results demonstrate what violation types are more difficult to agree upon and also which annotator tends to produce more annotations for particular tags in comparison with other annotator(s). This is the reason why Precision and Recall have been preferred to the Kappa coefficient.

5.4.1 Agreement for Annotations of Entity Mentions

The absolute counts for *entity mention* annotations for annotators 1 and 2 are shown in Table 5.3, in which the agreement scores using Precision, Recall and F-score for individual violation type of the *entity mention* tag for all pairs of annotators are also included.

In general, annotator 1 has more *entity mention* annotations than annotator 2. The biggest difference in counts is for *indefinite noun phrase with reference to previous mention* (INP+REF) and *pronoun with missing antecedent* (PRN+MISSA) violations, for which annotator 1 has twice as many annotations as annotator 2. *First mention without explanation* (FM-EXPL) and *definite noun phrase without reference to previous mention* (DNP-REF) violations represent the vast majority of the entity mention violations assigned.

	Counts		**Matches**	**Precision(1-2)**	**Precision (2-1)**	**F-score**
	1	**2**		**Recall (2-1)**	**Recall (1-2)**	
FM-EXPL	36	26	22	61.1	84.6	70.9
SM+EXPL	6	4	4	66.7	100.0	80.0
ACR-EXPL*	1	1	1	100.0	100.0	100.0
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
Total	**115**	**74**	**63**			
Average*				**54.5**	**90.4**	**67.5**

Table 5.2. Counts of *entity mention* annotations for annotators 1 and 2 and Precision, Recall and F-score for annotators 1-2 and 2-1 (numbers for annotations with * are excluded from calculation of average due to the low frequency)

In each pair of annotators, the first annotator is treated as the gold standard, for example in pair of annotators 1 and 2, annotator 1 is the gold standard and the annotations of annotator 2 are scored for Precision and Recall.

On average, annotators 1 and 2 and annotators 2 and 1 agree slightly above half the time. For annotators 1 and 2 F-score (67.5) is mainly affected by the lower precision (on average, Precision = 54.4), while for annotators 2 and 1 F-score is influenced by the lower recall (on average, Recall = 54.5).

Both pairs of annotators 1 and 2 and annotators 2 and 1 agree on *acronym without explanation* (ACR-EXPL) violation type, but this violation type has only one instance in the whole set. The reason for the full agreement could be the provision of the list of acronyms, which are supposed to have explanations in summaries (see Table 5.1).

In terms of annotations identified, agreement between annotators 1 and 2 is reached about half the time for most of annotations, except for two entity mention violation types: *indefinite noun phrase with reference to previous mention* (INP+RED) and *pronoun with missing antecedent* (PRN+MISLA), where Precision is below 50% for annotators 1 and 2 and Recall is below 50% for annotators 2 and 1.

Due to a very low frequency in the subset selected for the inter-annotator agreement, *acronym without explanation* (ACR-EXPL) and *pronoun with misleading antecedent* (PRN+MISLA) have been excluded from calculation of the average scores.

5.4.2 Agreement for Annotations of Relations between Entities

The results for the agreement on *relation between entities* annotations using Precision, Recall and F-score and also the absolute counts for individual annotations are shown in Table 5.4 for the pairs of annotators - 1 and 2, and 2 and 1.

On average, the agreement scores show high performance for both *link to first mention* (LINK2FM) and *link to previous mention* (LINK2PRVM) (F-score = 74.4), while the violation of *link to misleading antecedent* (LINK2MISLA) has been excluded from the calculation of the average score due to its low frequency in the set.

Table 5.4 also shows that the most frequent violation type is *link to previous mention* (LINK2PRVM), for which annotator 1 again has almost twice as many annotations as annotator 2.

	Counts **1**	 **2**	**Matches**	**Precision(1-2)** **Recall (2-1)**	**Precision (2-1)** **Recall (1-2)**	**F-score**
LINK2FM	6	4	4	66.7	100.0	80.0
LINK2PRVM	20	12	11	55.0	91.7	68.7
LINK2MISLA*	1	2	1	100.0	50.0	66.7
Total	**27**	**18**	**16**			
Average*				**60.9**	**95.9**	**74.4**

Table 5.3. Counts of *relation between entities* annotations for annotators 1 and 2 and Precision, Recall and F-score for annotators 1-2 and 2-1 (numbers for annotations with * are excluded from calculation of average due to the low frequency)

5.4.3 Agreement for Annotations of Clauses

Table 5.5 presents the counts for *clause* annotations and also the agreement scores, using Precision, Recall and F-score. Both annotators 1 and 2 have roughly the same amount of annotations for each individual violation type of the *clause* annotation tag.

On average, a high level of agreement has been reached for all violations of the *clause* tag (F-score 88.5). The violation types *incomplete sentence* (INCOMPLSN) and *inclusion of datelines* (INCLDATE) have particularly high agreement scores - 94.3 and 95.8, respectively, which means that annotators 1 and 2 agree on the number and spans of annotations in almost all cases, suggesting that it is fairly easy to determine what constitutes the ungrammaticality of clauses.

A very high rate of agreement has been reached for all pairs of annotators when annotating incomplete sentences, datelines and other ungrammatical forms, which suggests that such types of violations can be easily detected in text.

	Counts **1**	 **2**	**Matches**	**Precision(1-2)** **Recall (2-1)**	**Precision (2-1)** **Recall (1-2)**	**F-score**
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
Total	**96**	**97**	**87**			
Average				**89.3**	**87.7**	**88.5**

Table 5.4. Counts of *clause* annotations for annotators 1 and 2 and Precision, Recall and F-score for annotators 1-2 and 2-1

5.4.4 Agreement for Annotations of Relations between Clauses

Table 5.6 shows the counts for all individual types of the *relation between clauses* tag and also the agreement rates for annotators 1 and 2, and 2 and 1, using Precision, Recall and F-score. As shown in the table, the most frequent violation type of the *clause* tag is *redundant information* (REDUNINF) and a high agreement for creating links between clauses has been also reached for redundant information (F-score = 69.0).

No matches have been found for the less commonly occurring *relations between clauses*, such as *no semantic relatedness* (NOSEMREL) and *no discourse relation* (NODISREL), most probably due to their sparseness.

	Counts		**Matches**	**Precision(1-2)**	**Precision (2-1)**	**F-score**
	1	**2**		**Recall (2-1)**	**Recall (1-2)**	
NOSEMREL	4	1	0	0	0	0
REDUNINF	28	26	19	65.5	73.1	69.0
NODISREL	3	0	0	0	0	0
Total	**35**	**27**	**19**			
Average				**65.5**	**73.1**	**69.0**

Table 5.5. Counts of *relation between clauses* annotations for annotators 1 and 2 and Precision, Recall and F-score for annotators 1-2 and 2-1

5.4.5 Agreement for Annotations of Misleading Discourse Connectives

Table 5.7 presents the counts of *misleading discourse connectives* (MISLDISCON) for annotators 1 and 2 and also the agreement rates for both pairs of annotators, using Precision, Recall and F-score.

The misleading discourse connectives did not occur that often in the subset, only 6 misleading discourse connectives have been detected in this subset are ”but”, ”but”, ”but”, ”so”, ”so” and ”meanwhile”. Annotator 1 and 2 agreed on two occurrences of ”but”, the rest of the connectives have not been identified by annotator 2.

	Counts		**Matches**	**Precision(1-2)**	**Precision (2-1)**	**F-score**
	1	**2**		**Recall (2-1)**	**Recall (1-2)**	
MISLDISCON	**6**	**2**	**2**	**33.3**	**100.0**	**50.0**

Table 5.6. Counts of *misleading discourse connective* annotations for annotators 1 and 2 and Precision, Recall and F-score for annotators 1-2 and 2-1

5.5 Summary

This chapter includes the results from inspecting summaries for linguistic quality violations, presents the design of a scheme for annotating different violations of linguistic quality that may occur in automatically generated summaries. This chapter also describes the annotation process and reports on the results of the inter-annotator agreement.

The aim of the developed annotation scheme is to provide consistent and fine-grained annotations that capture information of the variety of linguistic violations. The agreement rates suggest that the scheme is well suited for this task resulting in high agreement between annotators across all annotations.

In addition, the inter-annotator agreement results demonstrate that some annotations are easier to be identified and categorized by different annotators. Though the annotation guidelines intended to make a considerable effort to enforce consistency of annotated spans, some agreement rates show that there may be some need to refine the guidelines in order to ensure total consistency amongst annotators.

Chapter 6

Corpus Data Analysis

In this chapter the corpus statistics are provided for the TAC and G-Flow datasets that have been annotated by the primary annotator only (annotator 1). The descriptions of both datasets can be found in Chapter 4. In addition, the comparison between these two datasets is provided.

6.1 Corpus Statistics

As shown in Table 6.1, the annotated corpus of the TAC dataset contains 1935 extracts of newspaper articles, while the annotated corpus of the entire G-Flow dataset contains only 50 extracts of newspaper articles.

	TAC	**G-Flow**
Total number of documents	1935	50

Table 6.1. Total number of the annotated TAC and G-Flow documents

Table 6.2 presents the absolute counts of all violation types across all annotation tags and also the average number of violations per document.

In total, 8001 annotations have been created for TAC, out of which 1541 annotations have no violation in linguistic quality and total of 37 annotations have been produced for G-Flow with 9 annotations have been assigned the violation type NONE violation, but which have been used for setting links between entities and clauses to indicate a certain type of violation of linguistic quality.

Violations	TAC		G-Flow	
	Count	Avg/doc	Count	Avg/doc
entity mention				
DNP-REF	958	0.50	3	0.06
FM-EXPL	792	0.41	6	0.12
INP+REF	430	0.22	1	0.02
PRN+MISSA	361	0.19	2	0.04
NONE	314	0.01	1	0
SM+EXPL	162	0.08	1	0.02
PRN+MISLA	27	0.01	0	0
ACR-EXPL	11	0.01	2	0.04
Total for entity mentions*	**2741**	**0.20**	**15**	**0.02**
relation between entities				
LINK2PRVM	436	0.23	1	0.02
LINK2FM	156	0.08	1	0.02
LINK2MISLA	27	0.01	0	0
Total for relations between entities	**619**	**0.11**	**2**	**0.01**
clause				
NONE	1227	0.63	8	0.16
INCOMPLSN	1044	0.54	0	0
OTHRUNGR	793	0.41	3	0.06
INCLDATE	412	0.21	3	0.06
Total for clauses*	**2249**	**0.39**	**6**	**0.04**
relation between clauses				
REDUNINF	504	0.26	3	0.06
NOSEMREL	142	0.07	0	0
NODISREL	91	0.05	1	0.02
Total for relations between clauses	**737**	**0.13**	**4**	**0.02**
misleading discourse connective				
MISLDISCON	114	0.06	1	0.02
Total for misleading discourse connectives	**114**	**0.06**	**1**	**0.02**
Total without NONE	**6460**	**0.19**	**28**	**0.03**
Total with NONE	**8001**	**0.21**	**37**	**0.04**

Table 6.2. Counts of all violation types identified in the TAC and G-Flow datasets (The total numbers with * exclude NONE)

6.1.1 Analysis for Entity Mention Annotations

As shown in Table 6.2, for TAC 2741 entity mention annotations are of certain violation type. Therefore, 314 entity mentions annotated with the violation type *none* have been excluded from the analysis as they do not represent any violation of linguistic quality. Table 6.2 also presents the frequency of the G-Flow *entity mention* violations with the total of 15 *entity mention* annotations. For TAC, the *entity mention* violation types *definite noun phrase without reference to previous mention* (DNP-REF), *first mention without explanation* (FM-EXPL), *indefinite noun phrase with reference to previous mention* (INP+REF) and *pronoun with missing antecedent* (PRN+MISSA) are the most frequently annotated types among all entity mention violations. For G-Flow, the most frequently occurring *entity mention* violation is the *first mention without explanation* (FM-EXPL) and *definite noun phrase without reference to previous mention* (DNP-REF). In comparison with G-Flow summaries, the TAC dataset on average contains more occurrences of such violations in their summaries: FM-EXPL for TAC (0.41) vs. FM-EXPL for G-Flow (0.12) and DNP-REF for TAC (0.50) vs. DNP-REF for G-Flow (0.06).

The high frequency of *definite noun phrase without reference to previous mention* (DNP-REF) and *indefinite noun phrase with reference to previous mention* (INP+REF) violation types provides evidence for the assertion that some systems that participated in the TAC 2011 have paid less attention to sentence ordering that organizes texts into coherent summaries.

First mention without explanation (FM-EXPL) and *pronoun with missing antecedent* (PRN+MISSA) being among the most dominant types of entity mention violations show yet another problem of multi-document summarization. Named entities represent the core information that must be present and properly described in summaries. Therefore, summaries containing people's full names and descriptions at their subsequent mentions, and using shorter ways, for example, last name only, to refer to people at first mentions make summaries less readable. Sentences being extracted without proper context often tend to have pronouns that lack antecedents, making it unclear for the reader who the pronouns are referring to.

Handling anaphora and references to named entities is a difficult task for certain summarization approaches that participated in TAC 2011 as on average the number of *pronouns with missing antecedents* (PRN+MISSA) for TAC is higher than that of G-Flow (0.19 vs. 0.04). The summarization systems that extract a sentence that starts with a pronoun more often than not ends up with a problem where a pronoun misses the antecedent. As a result, such violations affect the continuity and readability of sentences in the summary. Not selecting any sentences that starts with a pronoun could work well to maintain readability, but it could also mean that a lot of important sentences may be eliminated from sentence selection.

6.1.2 Analysis for Relation between Entities Annotations

Table 6.2 shows that in total, 619 relations have been created between pairs of entity mentions for TAC, while for G-Flow only 2 relations between entities have been identified.

Table 6.2 also contains information on the frequency for each of the violation types for the *relation between entities* annotation tag for both TAC and G-Flow datasets. The most frequent violation type for *relation between entities* annotations, by a significant margin, is the *link to previous mention* (LINK2PRVM), which is unsurprising given that *indefinite noun phrase with reference to previous mention* (INP+REF) described in section 6.1.1 is also a frequently occurring type of violation.

Since there was *no pronoun with misleading antecedent* (PRN+MISLA) detected in the course of annotation of the G-Flow dataset, therefore, *no link to misleading antecedent* (LINK2MISLA) has also been created.

6.1.3 Analysis for Clause Annotations

Table 6.2 shows that 2249 clauses of a certain violation type have been annotated for TAC. The total number of *clause* annotations for the TAC dataset is 3476, which excludes 1227 *clause* annotations that have been assigned the violation type *none* (NONE). For G-Flow the total number of clause annotations is 14 with 8 clauses assigned *none* (NONE) violation type, which have been excluded from the analysis as these 8 clauses are grammatically correct, and only used to create a link between each other to demonstrate a certain type of violated relation between clauses.

Coherent summaries are supposed to be well-organized and formed by using only complete sentences. As shown in Table 6.2 1044 *incomplete sentences* (INCOMPLSN) represent the most dominant ungrammaticality violation type in TAC (on average, 0.54), while no sentences in the G-Flow dataset have been detected as incomplete.

The truncation of the TAC summaries that is enforced due to the 100-word length limit causes such a high number of incomplete sentences in the TAC dataset. Though the G-Flow summaries are also around 100 words long, all sentences in the summaries, nevertheless, are complete.

An interesting point to note is that the next commonly occurring violation type in TAC is the *other ungrammatical form* (OTHRUNGR) with total of 793 annotations. The average number of various ungrammatical forms detected in the TAC dataset (0.41) is higher than that for the G-Flow summaries (0.06)

The reason for such a high number of ungrammatical errors in extracts could be due to the implemented sentence compression algorithms in some summarization tools that participated in TAC 2011 and that generated the final summaries that are weak from the linguistic point of view. Also during the compression process some important information tends to get lost making the summary less readable.

Such ungrammaticality issues as lack of punctuation, when a summary is like a bag of words, complete absence of articles and spaces between words, capitalization errors can be found in the TAC summaries. Since the average number of ungrammatical forms that has been found in the G-Flow dataset is very low, the G-Flow summaries are more fluent and readable as they have more accurate sentence breaks, presence of spaces between words, correct quotation mark matching and correct punctuation marks in sentences that definitely contribute to the improved linguistic quality of summaries.

On average the TAC dataset contains many more violations of grammaticality in clauses that the G-Flow dataset (0.39 vs. 0.04).

6.1.4 Analysis for Relation between Clauses Annotations

As shown in Table 6.2, the total of 737 links between clauses have been created for TAC, while for G-Flow only 4 annotations have been created for the *relation between clauses* annotation type.

For TAC, the *redundant information* (REDUNINF) violation type is the most frequently occurring violation type, by a significant margin, with 504 pairs of clauses that stand in a redundant relation with each other. For G-Flow the most frequent violation type is also the *redundant information* (REDUNINF). These results suggest that the amount of redundancy remains one of the main issues for the extractive multi-document summarization, where similar sentences tend to appear in multiple articles and the final summaries are formed by picking the same or similar sentences from multiple documents that form a document set.

It is undesirable for the coherent summary to have any repetitions of the same information, as the summaries that contain a lot of redundant information are considered to be weak from the point of view of linguistic quality. Detecting and removing redundancy for the summarization systems that participated in the TAC multi-document summarization task seems to be a challenging task as the amount of redundant information in the TAC dataset on average (0.26) is higher than that of the G-Flow dataset (0.06).

Removing redundant information by compressing or fusing the extracted sentences could be one of the possible solutions for the redundancy problem.

Another frequent violation type for TAC is *no semantic relatedness* (NOSEMREL) with

142 pairs of adjacent clauses that are semantically unrelated. For the G-Flow dataset no sets of adjacent sentences have been annotated as semantically unrelated.

6.1.5 Analysis for Misleading Discourse Connectives Annotations

Table 6.2 shows the total of 114 *misleading discourse connectives* (MISLDISCON) that have been detected in the course of annotation of the TAC dataset. Only one discourse connective has been annotated as misleading for the G-Flow dataset and, therefore, only one set of clauses have been annotated as having *no discourse relation* (NODISREL) (see section 6.1.4).

6.2 Correlation of Violation Types with TAC Readability Scores

In the TAC 2011 Guided Summarization Task the summaries of the participating systems have been evaluated with regard to readability which is defined as a combination of five aspects of linguistic quality: grammaticality, non-redundancy, referential clarity, focus and structure/coherence. Readability is assessed by making humans rate summaries based on the above set of linguistic criteria by manually assigning a score on a scale from 1 to 5, where 5 is assigned to summaries with the best readability, and 1 to summaries with poor readability (see section 4.2).

Since the readabilty factor is a combination of different aspects, the manual evaluation of summaries is inconclusive of what particular aspect of linguistic quality a human assessor focuses on when evaluating a summary. Obviously, assessors are unlikely to focus on a single aspect of linguistic quality exclusively while totally ignoring the rest (Pitler et al., 2010). Therefore, we hypothesize that assessors are likely to assign lower readability scores to summaries if particular violations of linguistic quality occur in the summary. Our study aims at finding out which factors are most disturbing regarding the perceived readability of a summary.

In order to identify what types of violations of linguistic quality influence rating, the Pearson's correlation coefficient between sums of certain violations per document and readability scores has been calculated by employing Pearson correlation using R[7].

Table 6.3 provides the correlation coefficients that show that there is a dependence between

[7] http://www.r-project.org/

certain violation types and readability scores that have been assigned to the summaries that contain these types of violations.

The negative correlations show that the readability score assigned to a summary tends to increase as the number of these violation types in a summary decreases. Therefore, the presence of these violation types in summaries influences humans when those assign readability scores to summaries.

Violation Type	**Pearson's r**
INCOMPLSN	-0.211
PRN+MISSA	-0.191
OTHRUNGR	-0.179
REDUNINFO	-0.159
NOSEMREL	-0.148
DNP-REF	-0.120
INCLDATE	-0.091
PRN+MISLA	-0.065

Table 6.3. Violation types and Pearson's r correlations (All these correlations are significant at p<0.01.)

The table contains only correlations at significance level p<0.01. Due to a low frequency in the dataset the negative correlations for the other violation types have not been included as violations that affect readability scores. As future work correlations between such violation types and readability scores could be tested using a higher number of occurrences of these types in the dataset.

As shown in Table 6.3 *incomplete sentences*, *pronouns with missing antecedents*, *other ungrammatical forms*, *redundant information* and *no semantic relatedness* between adjacent sentences influence readabilty scores, which means that these types of violations disrupt the fluency of summaries and make human assessors assign lower readability scores to such summaries.

6.3 Interaction of Content with Readability

Poor content coverage of the summary often makes humans assign lower readability scores even if no violations of linguistic quality have been found in that summary. Therefore, it is difficult to draw apart the content of the summary from its readability.

For instance, the TAC summary in (1) has nothing to annotate from the point of view of linguistic quality, but gets a very low readability score of 1. The reason for this could be the close relation between content and readability, so one could assume that content often influences the score that is assigned to evaluate readability.

(1) Indonesian navy ships Wednesday battled storms and gale force winds to rescue 13 survivors of a ferry disaster, with hundreds still missing five days after the ship sank in a storm off Java island. ”(Warship) KRI Hiu saved 13 survivors just south of Bawean island today (Wednesday),” Central Java navy commander Colonel Yan Simamora told AFP. Bawean is 350 kilometres (220 miles) northeast of the Javanese port of Semarang, the original destination of the ill-fated ferry.

Some summaries in the TAC dataset contained one or two minor violations to annotate that should not have a great effect on the readability score of the summary, but they still received low readability scores.

The TAC summary in (2) has also been assigned the readability score of 1 and where besides the entity mention *these corals* that has been annotated and assigned a violation type of *definite noun phrase without reference to previous mention* (DNP-REF) and the last sentence that is incomplete, no other violations of linguistic quality have been identified. So due to the close interaction of content with readability the summary could have received such a low readability score because the human assessor did not find the content of the summary covered sufficiently enough.

(2) ”This is the best possible decision to start getting the trade in these corals under some form of international control,” said Ernie Cooper, a coral trade expert from wildlife monitoring network TRAFFIC. Growing awareness of environmental issues means that the tourism industry has lately been a partner to conservation efforts in major reef areas. The Nature Conservancy’s Coral Triangle Center works at several sites in Indonesia, including the Raja Ampat Islands in Papua and Komodo National Park, a major protected marine area in the Lesser Sunda Islands. The reefs are rich in coral species and home to up to 1,000

6.4 Other Observations

In the course of manual analysis of annotated summaries other interesting observations have been made and are explained below.

Some TAC summaries like the one in (3) have many occurrences of when words lack spaces between them, and the content seems to be fully covered and this summary still gets a high readability score of 4. One can assume that at times no spacing between words or other ungrammatical forms do not have any influence on the readability, if the content has been properly covered.

(3) The United States said Friday it was unaware of any internationalinvolvement in a rocket attack on the U.S. Embassy in Greece, butthe investigation was in its early stages. It was the first major attack against a U.S. target in more thana decade, following the arrest of members of Greece's far-leftNovember 17 terrorist group. Public Order Minister Vyron Polydoras said Greece ”stronglycondemns” the attack on the heavily guarded building. November 17 carried out a similar rocket attack against the U.S.embassy in 1996, causing minor damage and no injuries. Greek police said a far-left group, Revolutionary Struggle, hadclaimed responsibility for the

A great amount of automatically generated summaries appear to be identical, meaning that summarization systems, especially belonging to one and the same team (like systems with IDs 39 and 40), tend to produce the same summaries. The manual analysis of the readability scores of some pairs of identical summaries shows that such summaries often get assigned different readability scores. This observation suggests that assigning scores by human assessors is not always consistent.

Example (4) contains the TAC summary that appeared in the dataset twice, but with two different readability scores 3 and 1.

(4) On Monday morning, Charles Carl Roberts IV entered the West Nickel Mines Amish School in Lancaster County and shot 10 girls, killing five. The gunman, Charles Carl Roberts IV, 32, a truck driver from the town of Bart, apparently killed himself, state police Commissioner Jeffrey B. Miller said. NICKEL MINES, Pennsylvania 2006-10-04 22:50:42 UTC Two relatives of the man who attacked Amish school girls said Monday they were not molested by him 20 years ago, as he had claimed in a phone call to his wife during the siege. Miller confirmed three dead at the scene.

6.5 Summary

This chapter presents the statistical data for the annotated TAC and G-Flow datasets and compares the amount of different types of violations of linguistic quality encountered in summaries in each of the datasets. The comparison provides evidence that the G-Flow system, indeed, aims at generating coherent summaries as on average the amount of violations is higher for the TAC dataset (0.19) than that for the G-Flow set (0.03).

Measuring the correlations between the sums of different violations per summary and readabililty scores assigned to the summaries by humans showed that certain types of violations of linguistic quality influence the readability scores.

During the annotation process some interesting observations have been made. One of such observations confirms the hypothesis that there is a close interaction between the perceived content and the readability of the summary as some summaries get assigned a low readabilty score even if these summaries contain no violations of linguistic quality. Therefore, humans find it hard to measure the readability of the summary without taking into account the content of that summary. As a result, summaries with poor content coverage that contain no violations of linguistic quality still tend to receive low readability scores.

Chapter 7

Conclusion

7.1 Results

In the course of this project, the design of the scheme for annotating events that represent various violations of linguistic quality in automatically generated summaries has been presented and a corpus of summaries annotated for a fixed set of linguistic quality violations has been produced.

The results suggest that the annotation scheme is well suited for detecting and analyzing the violations of linguistic quality. Most of the annotations described in the scheme were assigned within the TAC and G-Flow datasets a sufficient number of times and provide evidence to their potential usefulness. Therefore, the annotated corpus could serve as a test set in evaluation of linguistic quality of automatically generated summaries.

In measuring the agreement between two annotators, the standard measure of Precision, Recall and F-score have been used. The agreement rates suggest that the annotation quality of the data is high for all annotation tags. Some violation types turned out to have a low frequency in the subset, so agreement could not be computed reliably for such types of annotations.

The results of the comparison of automatically generated summaries by the TAC and G-Flow datasets demonstrated that even the best content selection approaches used by the systems that participated in the TAC 2011 cannot achieve the same linguistic quality as the summaries generated by the G-Flow system that has been designed specifically for producing coherent summaries. Therefore, the linguistic quality of summaries could be improved if the summarization systems focused not solely on content coverage, but also on incorporating coherence in sentence extraction for producing more readable summaries.

In the course of this project, a number of violation types have been identified as the ones that tend to affect the readability scores assigned by humans. Therefore, the presence of such violations in summaries makes human assessors rate these summaries with lower readability scores.

7.2 Future Work

In the future, this work could be extended by applying a slightly modified annotation scheme to the TAC dataset collection used for the "Update Summarization Task" that contains summaries which include the information on the same topics, but with the assumption that the reader is already aware of the events described in the summary.

The annotation scheme suggested in this project could also be applied to violations of linguistic quality that occur in abstracts, where the content of the summaries is somewhat different from that used to form extracts.

Additionally, the annotated corpus could be used to define possible patterns according to which humans assign particular readability scores which could further contribute to the development of automatic evaluation metrics of linguistic quality of summaries.

Appendix A

List of Acronyms

Table A.1 contains the list of acronyms which have explanations in original documents in the TAC dataset. These acronyms are not well known to an average reader and therefore, one would also expect to have explanations of these acronyms in summaries.

Acronym	Explanation
RSF	Reporters Without Borders (possibly a translation)
WFP	United Nations World Food Program
CHD	Congenital heart disease
AHA	American Heart Association
ELA	People's Revolutionary Struggle (possibly a translation)
KNAPK	Greenland Hunters and Fishers Association (possibly a translation)
UVB	Ultraviolet B
CITES	Convention on International Trade in Endangered Species
BNP	Brain-type Natriuretic Peptide
JCI	Joy Contractors Inc.
NPC	National People's Congress
UC	University of California
CA	Chief Adviser
FFWC	Flood Forecasting and Warning Center

Table A.1. Acronyms with explanation in original documents

Appendix B

List of Files used for Inter-Annotator Agreement

B.1 A subset of the TAC dataset

1. D1101-A.peers.csv_2.xml
2. D1101-A.peers.csv_9.xml
3. D1101-A.peers.csv_28.xml
4. D1101-A.peers.csv_39.xml
5. D1102-A.peers.csv_2.xml
6. D1102-A.peers.csv_37.xml
7. D1102-A.peers.csv_41.xml
8. D1103-A.peers.csv_9.xml
9. D1103-A.peers.csv_11.xml
10. D1103-A.peers.csv_14.xml
11. D1103-A.peers.csv_39.xml
12. D1104-A.peers.csv_14.xml
13. D1104-A.peers.csv_26.xml
14. D1104-A.peers.csv_42.xml
15. D1105-A.peers.csv_25.xml
16. D1105-A.peers.csv_31.xml
17. D1105-A.peers.csv_33.xml
18. D1105-A.peers.csv_43.xml
19. D1106-A.peers.csv_12.xml
20. D1106-A.peers.csv_25.xml
21. D1106-A.peers.csv_34.xml
22. D1107-A.peers.csv_14.xml
23. D1107-A.peers.csv_16.xml
24. D1107-A.peers.csv_31.xml
25. D1107-A.peers.csv_37.xml
26. D1108-A.peers.csv_26.xml
27. D1108-A.peers.csv_40.xml
28. D1109-A.peers.csv_24.xml
29. D1109-A.peers.csv_37.xml
30. D1109-A.peers.csv_43.xml
31. D1110-A.peers.csv_16.xml

32. D1110-A.peers.csv_40.xml
33. D1111-A.peers.csv_4.xml
34. D1111-A.peers.csv_32.xml
35. D1111-A.peers.csv_43.xml
36. D1112-A.peers.csv_21.xml
37. D1112-A.peers.csv_22.xml
38. D1113-A.peers.csv_24.xml
39. D1113-A.peers.csv_35.xml
40. D1113-A.peers.csv_48.xml
41. D1114-A.peers.csv_13.xml
42. D1114-A.peers.csv_21.xml
43. D1115-A.peers.csv_3.xml
44. D1115-A.peers.csv_20.xml
45. D1115-A.peers.csv_48.xml
46. D1116-A.peers.csv_35.xml
47. D1116-A.peers.csv_47.xml
48. D1117-A.peers.csv_13.xml
49. D1117-A.peers.csv_18.xml
50. D1118-A.peers.csv_20.xml
51. D1119-A.peers.csv_8.xml
52. D1119-A.peers.csv_19.xml
53. D1120-A.peers.csv_45.xml
54. D1121-A.peers.csv_23.xml
55. D1121-A.peers.csv_34.xml
56. D1122-A.peers.csv_34.xml
57. D1123-A.peers.csv_8.xml
58. D1123-A.peers.csv_23.xml
59. D1124-A.peers.csv_10.xml
60. D1125-A.peers.csv_1.xml
61. D1125-A.peers.csv_19.xml
62. D1126-A.peers.csv_28.xml
63. D1127-A.peers.csv_11.xml
64. D1127-A.peers.csv_20.xml
65. D1128-A.peers.csv_5.xml
66. D1129-A.peers.csv_25.xml
67. D1130-A.peers.csv_42.xml
68. D1131-A.peers.csv_13.xml
69. D1131-A.peers.csv_35.xml
70. D1132-A.peers.csv_33.xml
71. D1132-A.peers.csv_48.xml
72. D1133-A.peers.csv_12.xml
73. D1133-A.peers.csv_35.xml
74. D1134-A.peers.csv_6.xml
75. D1134-A.peers.csv_41.xml
76. D1135-A.peers.csv_39.xml
77. D1135-A.peers.csv_44.xml
78. D1136-A.peers.csv_14.xml
79. D1136-A.peers.csv_17.xml
80. D1137-A.peers.csv_21.xml
81. D1137-A.peers.csv_22.xml
82. D1138-A.peers.csv_24.xml
83. D1138-A.peers.csv_25.xml
84. D1139-A.peers.csv_24.xml

85. D1139-A.peers.csv_32.xml
86. D1140-A.peers.csv_31.xml
87. D1140-A.peers.csv_44.xml
88. D1141-A.peers.csv_22.xml
89. D1141-A.peers.csv_23.xml
90. D1142-A.peers.csv_12.xml
91. D1142-A.peers.csv_34.xml
92. D1143-A.peers.csv_17.xml
93. D1143-A.peers.csv_29.xml
94. D1144-A.peers.csv_10.xml
95. D1144-A.peers.csv_37.xml

B.2 A subset of the G-Flow dataset

96. D30001.M.100.T.xml
97. D30015.M.100.T.xml
98. D30020.M.100.T.xml
99. D30045.M.100.T.xml
100. D31050.M.100.T.xml

Appendix C

Inter-Annotator Agreement for annotators 1 and 3 and annotators 2 and 3

C.1 Annotation of Entity Mentions

Entity mention	**1**	**2**	**3**
FM-EXPL	36	26	33
SM+EXPL	6	4	2
ACR-EXPL	1	1	1
DNP-REF	34	23	20
INP+REF	19	9	5
PRN+MISSA	18	9	10
PRN+MISLA	1	2	1
Total	**115**	**74**	**72**

Table C.1. Counts of *entity mention* annotations for annotators 1, 2 and 3

	Counts		**Matches**	**Precision(1-3)**	**Precision (3-1)**	**F-score**
	1	**3**		**Recall (3-1)**	**Recall (1-3)**	
FM-EXPL	36	33	19	52.8	57.6	55.1
SM+EXPL	6	2	2	33.3	100.0	50.0
ACR-EXPL*	1	1	1	100.0	100.0	100.0
DNP-REF	34	20	11	32.4	55.0	40.7
INP+REF	19	5	5	26.3	100.0	41.7
PRN+MISSA	18	10	10	55.6	100.0	71.4
PRN+MISLA*	1	1	0	0	0	0
Total	**115**	**72**	**48**			
Average*				**40.1**	**82.5**	**51.8**

Table C.2. Counts of *entity mention* annotations for annotators 1 and 3 and Precision, Recall, F-score for annotators 1-3 and 3-1 (numbers for annotations with * are excluded from calculation of average due to the low frequency)

	Counts		**Matches**	**Precision(2-3)**	**Precision (3-2)**	**F-score**
	2	**3**		**Recall (3-2)**	**Recall (2-3)**	
FM-EXPL	26	33	15	57.7	45.5	50.8
SM+EXPL	4	2	1	25.0	50.0	33.3
ACR-EXPL*	1	1	1	100.0	100.0	100.0
DNP-REF	23	20	10	43.5	50.0	46.5
INP+REF	9	5	5	55.6	100.0	71.4
PRN+MISSA	9	10	7	77.8	70.0	73.7
PRN+MISLA*	2	1	0	0	0	0
Total	**74**	**72**	39			
Average*				**51.9**	**63.1**	**55.1**

Table C.3. Counts of *entity mention* annotations for annotators 2 and 3 and Precision, Recall and F-score for annotators 2-3 and 3-2 (numbers for annotations with * are excluded from calculation of average due to the low frequency)

C.2 Annotation of Relations between Entities

Relation between entities	1	2	3
LINK2FM	6	4	15
LINK2PRVM	20	12	7
LINK2MISLA	1	2	0
Total	**27**	**18**	**22**

Table C.4. Counts of *relation between entities* annotations for annotator 1, 2 and 3

	Counts		Matches	Precision(1-3)	Precision (3-1)	F-score
	1	3		Recall (3-1)	Recall (1-3)	
LINK2FM	6	15	4	66.7	26.7	38.1
LINK2PRVM	20	7	4	20.0	57.1	29.6
LINK2MISLA*	1	0	0	0	0	0
Total	**27**	**22**	**8**			
Average*				**43.4**	**41.9**	**33.9**

Table C.5. Counts of *relation between entities* annotations for annotators 1 and 3 and Precision, Recall and F-score for annotators 1-3 and 3-1 (numbers for annotations with * are excluded from calculation of average due to the low frequency)

	Counts		Matches	Precision(2-3)	Precision (3-2)	F-score
	2	3		Recall (3-2)	Recall (2-3)	
LINK2FM	4	15	2	50.0	13.3	21.1
LINK2PRVM	12	7	4	33.3	57.1	42.1
LINK2MISLA*	2	0	0	0	0	0
Total	**18**	**22**	**6**			
Average*				**41.7**	**35.2**	**31.6**

Table C.6. Counts for *relation between entities* annotations for annotators 2 and 3 and Precision, Recall and F-score for annotators 2-3 and 3-2 (numbers for annotations with * are excluded from calculation of average due to the low frequency)

C.3 Annotation of Clauses

Clause	**1**	**2**	**3**
INCOMPLSN	43	44	46
INCLDATE	24	24	23
OTHRUNGR	29	29	42
Total	**96**	**97**	**111**

Table C.7. Counts of *clause* annotations for annotator 1, 2 and 3

	Counts		**Matches**	**Precision(1-3)**	**Precision (3-1)**	**F-score**
	1	**3**		**Recall (3-1)**	**Recall (1-3)**	
INCOMPLSN	43	46	39	90.7	84.8	87.6
INCLDATE	24	23	23	95.8	100.0	97.9
OTHRUNGR	29	42	30	85.7	69.8	76.9
Total	**96**	**111**	**92**			
Average				**90.7**	**84.9**	**87.5**

Table C.8. Precision and Recall for *clause* annotations for annotators 1-3 and 3-1

	Counts		**Matches**	**Precision(2-3)**	**Precision (3-2)**	**F-score**
	2	**3**		**Recall (3-2)**	**Recall (2-3)**	
INCOMPLSN	44	46	38	86.4	82.6	84.4
INCLDATE	24	23	23	95.8	100.0	97.9
OTHRUNGR	29	42	23	65.7	53.5	59.0
Total	**97**	**111**	**84**			
Average				**82.6**	**78.7**	**80.4**

Table C.9. Counts of *clause* annotations for annotators 2 and 3 and Precision, Recall and F-score for annotators 2-3 and 3-2

C.4 Annotation of Relations between Clauses

Relation between clauses	1	2	3
NOSEMREL	4	1	0
REDUNINF	28	26	23
NODISREL	3	0	0
Total	**35**	**27**	**23**

Table C.10. Counts of *relation between clauses* annotations for annotator 1, 2 and 3

	Counts		**Matches**	**Precision(1-3)**	**Precision (3-1)**	**F-score**
	1	**3**		**Recall (3-1)**	**Recall (1-3)**	
NOSELREL	4	0	0	0	0	0
REDUNINF	28	23	14	50.0	60.9	54.9
NODISREL	3	0	0	0	0	0
Total	**35**	**23**	**14**			
Average				**50.0**	**60.9**	**54.9**

Table C.11. Counts of *relation between clauses* annotations for annotators 1 and 3 and Precision, Recall and F-score for annotators 1-3 and 3-1

	Counts		**Matches**	**Precision(2-3)**	**Precision (3-2)**	**F-score**
	2	**3**		**Recall (3-2)**	**Recall (2-3)**	
NOSEMREL	1	0	0	0	0	0
REDUNINF	26	23	15	57.7	62.5	60.0
NODISREL	0	0	0	0	0	0
Total	**27**	**23**	**15**			
Average				**57.7**	**62.5**	**60.0**

Table C.12. Counts of *relation between clauses* annotations for annotators 2 and 3 and Precision, Recall and F-score for annotators 2-3 and 3-2

C.5 Annotation of Misleading Discourse Connectives

misleading discourse connective	1	2	3
MISLDISCON	6	2	5

Table C.13. Counts of *misleading discourse connectives* for annotators 1, 2 and 3

	Counts		**Matches**	**Precision(1-3)**	**Precision (3-1)**	**F-score**
	1	**3**		**Recall (3-1)**	**Recall (1-3)**	
MISLDISCON	**6**	**5**	**2**	**33.3**	**40.0**	**36.4**

Table C.14. Counts of *misleading discourse connective* annotations for annotators 1 and 3 and Precision, Recall and F-score for annotators 1-3 and 3-1

	Counts		**Matches**	**Precision(2-3)**	**Precision (3-2)**	**F-score**
	2	**3**		**Recall (3-2)**	**Recall (2-3)**	
MISLDISCON	**2**	**5**	**2**	**100.0**	**40.0**	**57.1**

Table C.15. Counts of *misleading discourse connective* annotations for annotators 2 and 3 and Precision, Recall and F-score for annotators 2-3 and 3-2

Bibliography

Bakawid, A. and Oussalah, M. (2011). Using SSM for Enhansing Summarization. In *TAC 2011 Workshop Proceedings*.

Barrera, A., Verma, R. M., and Vincent, R. (2011). SemQuest: University of Houston's Semantics-based Question Answering System. In *TAC 2011 Workshop Proceedings*.

Barzilay, R. and Lapata, M. (2005). Modeling Local Coherence: An Entity-based Approach. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 141—-148.

Baxendale, P. B. (1958). Machine-made index for technical literature: an experiment. In *IBM Journal of Research Development*, pages 354–361.

Chali, Y., Hasan, S. A., Imam, K., and Subramanian, S. (2011). UofL at TAC 2011 Guided Summarization Task. In *TAC 2011 Workshop Proceedings*.

Christensen, J., Mausam, Soderland, S., and Etzioni, O. (2013). Towards Coherent Multi-Document Summarization. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2013)*, Atlanta, Georgia.

Conroy, J. M., Schlesinger, J. D., Kubina, J., Rankel, P. A., and O'Leary, D. P. (2011). CLASSY 2011 at TAC: Guided and Multi-lingual Summaries and Evaluation Metrics. In *TAC 2011 Workshop Proceedings*.

Das, D. and Martins, A. F. (2007). A Survey on Automatic Text Summarization.

Das, P. and Srihari, R. (2011). Global and Local Models for Multi-Document Summarization. In *TAC 2011 Workshop Proceedings*.

Du, P., Yuan, J., Lin, X., Zhang, J., Guo, J., and Cheng, X. (2011). Decayed DivRank for Guided Summarization. In *TAC 2011 Workshop Proceedings*.

Edmundson, H. P. (1969). New methods in automatic extracting. In *Journal of the Association for Computing Machinery*, pages 264–285.

Elsner, M. and Charniak, E. (2008). Coreference-inspired Coherence Modeling. In *Proceedings of ACL – HLT'08*.

Hahn, U. and Mani, I. (2000). The Challenges of Automatic Summarization. pages 29–36.

He, R., Fu, K., and Zhou, X. (2011). Category oriented Extractive Content Selection for Guided Summarization. In *TAC 2011 Workshop Proceedings*.

Hovy, E. (2005). Automated Text Summarization. In *The Oxford Handbook of Computational Linguistics*, pages 583—-598.

Hu, P. and Ji, D. (2011). WHUSUM Participation at TAC 2011 Guided Summarization Track. In *TAC 2011 Workshop Proceedings*.

Jurafsky, D. and Martin, J. H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

Kennedy, A., Kazantseva, A., Mohammad, S., Copeck, T., Inkpen, D., and Szpakowicz, S. (2011). Getting Emotional About News. In *TAC 2011 Workshop Proceedings*.

Klassen, P. (2011). University of Washington at TAC 2011. In *TAC 2011 Workshop Proceedings*.

Kumar, N., Srinathan, K., and Varma, V. (2011). Using Unsupervised System with least linguistic feutures for TAC-AESOP Task. In *TAC 2011 Workshop Proceedings*.

Lapata, M. and Barzilay, R. (2005). Automatic Evaluation of Text Coherence: Models and Representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1085–1090.

Li, H., Hu, Y., Li, Z., Wan, X., and Xiao, J. (2011a). PKUTM participation in TAC2011. In *TAC 2011 Workshop Proceedings*.

Li, S., Song, T., and Wang, X. (2011b). TAC 2011 Guided Summarization of ICL. In *TAC 2011 Workshop Proceedings*.

Lin, C.-Y. (2004). Rouge: a package for automatic evaluation of summaries. In *Proceedings of the ACL Text Summarization Workshop on Text Summarization Branches Out. ACL*, pages 25–26.

Lin, C.-Y. and Hovy, E. (1997). Identifying Topics by Position. In *Proceedings of the Fifth conference on Applied Natural Language Processing*, pages 283–290.

Lin, C.-Y. and Hovy, E. (2002). Manual and automatic evaluation of summaries. In *Proceedings of the Workshop on Automatic Summarization, post conference workshop of ACL 2002*, pages 45–51.

Liu, H., Liu, P., Heng, W., and Li, L. (2011). The CIST Summarization System at TAC 2011. In *TAC 2011 Workshop Proceedings*.

Louis, A. and Nenkova, A. (2009). Automatically Evaluating Content Selection in Summarization without Human Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Luhn, H. P. (1958). The automatic creation of literature abstracts. In *IBM Journal of Research Development*, pages 159–165.

Makino, T., Takamura, H., and Okumura, M. (2011). Balanced Coverage of Aspects for Text Summarization. In *TAC 2011 Workshop Proceedings*.

Mani, I. (2001). Summarization Evaluation: An Overview. In *Proceedings of the NAACL 2001 Workshop on Automatic Summarization*.

Mason, R. and Charniak, E. (2011). Extractive multi-document summaries should explicitly not contain document-specific content. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, pages 49–54.

Nenkova, A. (2005). Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference. In *American Association for Artificial Intelligence*, pages 1436–1441.

Nenkova, A., Chae, J., Louis, A., and Pitler, E. (2010). Structural Features for Predicting the Linguistic Quality of Text - Applications to Machine Translation, Automatic Summarization and Human-Authored Text. In *Empirical Methods in Natural Language Generation*, pages 222–241.

Nenkova, A. and McKeown, K. (2003). References to Named Entities: a Corpus Study. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

Nenkova, A. and McKeown, K. (2011). Automatic Summarization. pages 103–233.

Nenkova, A. and Passonneau, R. (2004). Evaluating Content Selection in Summarization: The Pyramid Method. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Nenkova, A., Passonneau, R., and Mckeown, K. (2007). The pyramid method: incorporating human content selection variation in summarization evaluation. In *ACM Transactions on Speech and Language Processing*.

Ng, J.-P., Bysani, P., Lin, Z., Kan, M.-Y., and Tan, C.-L. (2011). SWING: Exploiting Category-Specific Information for Guided Summarization. In *TAC 2011 Workshop Proceedings*.

Pitler, E., Louis, A., and Nenkova, A. (2010). Automatic Evaluation of Linguistic Quality in Multi-Document Summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.*

Pitler, E. and Nenkova, A. (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.

Steinberger, J., Kabadjov, M., Steinberger, R., Tanev, H., Turchi, M., and Zavarella, V. (2011). JRC's Participation at TAC 2011: Guided and Multilingual Summarization Tasks. In *TAC 2011 Workshop Proceedings.*

Stubbs, A. (2011). MAE and MAI: Lightweight Annotation and Adjudication Tools. In *2011 Proceedings of the Linguistic Annotation Workshop V, Association of Computational Linguistics*, Portland, Oregon.

Vadlapudi, R. and Katragadda, R. (2010). On automated evaluation of readability of summaries: capturing grammaticality, focus, structure and coherence. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 7–12.

Varma, V., Kovelamudi, S., Gupta, J., Priyatam, N., Sood, A., Jain, H., Mogadala, A., and Vaddepally, S. R. (2011). III Hyderabad in Summarization and Knowledge Base Population at TAC2011. In *TAC 2011 Workshop Proceedings.*

Zhang, J., Yao, C., Ding, X., Li, Z., Xu, W., Chen, G., and Guo, J. (2011a). PRIS at TAC2011 Guided Summarization Track. In *TAC 2011 Workshop Proceedings.*

Zhang, R., Ouyang, Y., and Li, W. (2011b). Guided Summarizatuon with Aspect Recognition. In *TAC 2011 Workshop Proceedings.*