Text Summarization Of Multiple Documents for Enterprise Specific Using Custom Named Entity Recognition

Revan Datta Parimi

Final Thesis Report

June 2024

# **TABLE OF CONTENTS**

acknowledgements 4

abstract 4

LIST OF TABLES 4

LIST OF FIGURES 4

LIST OF ABBREVIATIONS 4

CHAPTER 1: INTRODUCTION 4

1.1 Background 4

1.2 Problem Statement & Related Works 5

1.3 Aim and Objectives 6

1.4 Research Questions 6

1.5 Scope of the Study 6

1.6 Significance of Study 7

CHAPTER 2: LITERATURE REVIEW 8

2.1 Introduction 8

2.1.1 History of Text Summarization 8

2.1.2 Applications of Text Summarization 8

2.1.3 Recommendations of Text Summarization 8

2.2 Text Summarization Fundamentals 8

2.2.1 Extractive vs Abstractive Methods 8

2.2.2 Evaluation Metrics 8

2.2.3 Key Challenges 9

2.3 Transfer Learning for Text Summarization 9

2.3.1 Domain Adaptation Techniques 9

2.3.2 Multi-Task Techniques 9

2.4 Customizing Language Models 10

2.4.1 Vocabulary Integration Approaches 10

2.4.2 Entity Recognition Extensions 10

2.5 Incorporating Domain Knowledge 10

2.5.1 Knowledge Bases for NLP 10

2.5.2 Rule Injection Methods 10

2.6 Summarization for Long Documents 11

2.6.1 Hierarchical Models 11

2.6.2 Salience Estimation 11

CHAPTER 3: RESEARCH METHODOLOGY 11

3.1 Data Selection 11

3.1.1 Purpose 11

3.1.2 Approach 11

3.1.3 Criteria 11

3.2 Pre-processing 12

3.2.1 Tockenization 12

3.2.2 Selective Stop word Removal 12

3.2.3 Lemmatization and Compound Word Handling 12

3.2.4 Part-of-Speech Tagging and named Entity Recognition 12

3.2.5 Code Extraction and Formatting 12

3.3 Data Engineering 12

3.3.1 Word Embeddings 12

3.3.2 Feature Engineering 13

3.3.3 Data Augmentation 13

3.4 Interactive Visual Analytics 13

3.4.1 Semantic Similarity Visualization 13

3.4.2 Topic Modelling and Clustering 13

3.4.3 Code Exploration 13

3.4.4 Collaborative Annotation 14

3.5 Interpretation and Evaluation 14

3.5.1 Precise, Recall, and F1-Score 14

3.5.2 Coherence and Readability Evaluation 14

3.5.3 Knowledge Retention Assessment 14

3.5.4 Error Analysis and Feedback Evaluation 15

CHAPTER 4: ANALYSIS 11

4.1 Introduction 11

4.2 Dataset Description 11

4.3 Exploratory Data Analysis 12

4.3.1 To 12

4.3.2 Sele 12

4.4 Data Visualization 12

4.4.1 W 12

4.4.2 F 13

4.4.3 D 13

4.5 Summary 13

CHAPTER 5: RESULTS AND DISCUSSIONS 11

5.1 Introduction 11

5.2 Interpretation of Visualizations 11

5.3 Evaluation of Sampling Methods and Results 12

5.3.1 To 12

5.3.2 Sele 12

5.4 Testing of Validation Dataset 12

5.4.1 W 12

5.4.2 F 13

5.4.3 D 13

5.5 Summary 13

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS 151

6.1 Introduction 151

6.2 Discussion and Conclusion 151

6.3 Contribution to Knowledge 153

6.4 Future Recommendations 153

REFERENCES 15

APPENDIX A: RESEARCH PROPOSAL 15

APPENDIX B: ETHIC FORMS 16

APPENDIX C: RESEARCH PAPERS 16

**Acknowledgments**

I would like to express my sincere gratitude to my supervisor, Dr. Sabyasachi Parida, for his continuous support, patience, and guidance throughout the course of my research. His insightful feedback and encouragement have been invaluable.

I am also grateful to the members of my thesis committee, Dr. Ahmed Kaky and Anil Vuppala, for their constructive sessions and suggestions, which significantly enhanced the quality of this work.

I extend my thanks to John Moore Liver Pool University for their support, which made this research possible. Additionally, I acknowledge the resources and support provided by the Department of Machine Learning and Artificial Intelligence at IIIT Bangalore.

I would like to thank my colleagues and friends in the research lab for the collaboration, and the stimulating discussions that greatly contributed to my research.

A special thanks to my family, particularly my parents and my partner, for their unwavering support and understanding during the challenging moments of this journey. Their love and encouragement have been my source of strength.

Lastly, I would like to acknowledge the administrative staff and professors at John Moore Liver Pool University for their prompt assistance and access to necessary resources.

**Abstract**

This thesis addresses the challenge of efficiently summarizing lengthy enterprise documents, such as financial reports and project plans, which contain valuable insights but are often too complex and time-consuming to analyze manually. Existing text summarization methods trained on open-domain datasets struggle with the unique vocabulary and specific entities found in enterprise content. This research develops customized named entity recognition (NER) and summarization techniques tailored to enterprise-specific documents.

The study begins with a comprehensive literature review, exploring the differences between extractive and abstractive summarization methods, and the challenges posed by domain-specific vocabulary. It then delves into the research methodology, including data selection and pre-processing techniques, and the customization of language models using SpaCy to handle enterprise terminology.

Through a series of experiments, the research evaluates the effectiveness of the proposed methods in summarizing enterprise documents. The results demonstrate significant improvements in capturing salient information and generating coherent summaries, thereby facilitating knowledge management and onboarding processes within enterprises.

The findings contribute to the field of natural language processing by highlighting the importance of domain adaptation and the integration of domain knowledge into summarization models. The study concludes with recommendations for future research, including the exploration of hierarchical summarization models and the integration of structured knowledge bases.

**LIST OF TABLES**

Table 1: SpaCy, Core NLP and NLTK (statistically)

|  |  |  |  |
| --- | --- | --- | --- |
| Package | Precision | Recall | F1 Score |
| SpaCy | 0.72 | 0.65 | 0.69 |
| Core NLP | 0.79 | 0.73 | 0.76 |
| NLTK | 0.51 | 0.65 | 0.58 |

Table 2: SpaCy, Core NLP, and NLTK (grammatically)

|  |  |  |
| --- | --- | --- |
| Package | Tokenization (milliseconds) | Tag (milliseconds) |
| SpaCy | 2 | 10 |
| Core NLP | 4 | 443 |
| NLTK | 0.2 | 1 |

Table 3: SpaCy vs NLTK

|  |  |  |
| --- | --- | --- |
| Models | Previous | Proposed |
| Feature | NLTK | SpaCy |
| Classifier | Yes | Yes |
| Topic Modelling | No | Yes |
| Vectorization | No | Yes |
| Tokenization | Yes | Yes |
| Parsing | Yes | Yes |
| TF-IDF | No | Yes |

**LIST OF FIGURES**

Figure 1: Flow Diagram



Figure 2: SpaCy Processing Pipeline



**LIST OF ABBREVIATIONS**

1. NLP (Natural Language Processing)
2. NER (Named Entity Recognition)
3. POS (Part-of-Speech)
4. TF-IDF (Term Frequency-Inverse Document Frequency)
5. SMEs (Subject Matter Experts)
6. LDA (Latent Dirichlet Allocation)
7. t-SNE (t-Distributed Stochastic Neighbor Embedding)
8. MMR (Maximal Marginal Relevance)
9. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
10. METEOR (Metric for Evaluation of Translation with Explicit ORdering)
11. DUC (Document Understanding Conference)
12. CNN (Convolutional Neural Network)
13. LSTM (Long Short-Term Memory)
14. QA (Question Answering)
15. NLTK (Natural Language Toolkit)
16. **Introduction**
    1. **Background**

Enterprise documents like financial reports, legal contracts, project plans, etc. contain valuable insights and key details related to business operations and decisions. However, these documents tend to be much longer compared to generic corpora. Manually analyzing such long enterprise documents is incredibly time-consuming and labor-intensive.

Text summarization methods can help by automatically identifying and extracting the most important information from enterprise documents and generating concise overviews. This provides analysts and decision-makers quick access to the key details without needing to read the full documents. However, standard text summarization techniques often perform poorly on enterprise documents containing unique industry-specific vocabulary and entities not found in their training data.

For instance, in credit rating agencies, there are large internal repositories of documents, wikis, and knowledge transfer content developed over many years that are critical for onboarding new analysts. But reading and absorbing all the material is infeasible. Text summarization tailored to such content could help accelerate understanding.

Likewise, many large enterprises have company-specific terminologies and named entities which pose challenges for off-the-shelf natural language processing tools. There is a need for customized techniques.

Past research has developed generalized text summarization methods using news and Wikipedia articles. However, these fail to handle the unique vocabulary and semantics of proprietary enterprise content. There has been limited work on adapting summarization for industry-specific documents.

This research aims to address that gap by developing tailored NER and summarization techniques to distill enterprise documents containing custom vocabulary and details not found in open datasets. The methods could help enterprises efficiently process large internal document collections without manual effort.

* 1. **Problem Statement and Related Works**
     1. **Problem Statement**

Enterprise documents like financial reports, project plans, and internal memos contain valuable company-specific details and knowledge. However, these documents tend to be much longer and more complex compared to general domain corpora. Manually reviewing all the content is incredibly time-consuming for employees and new hires.

Existing text summarization systems are trained and evaluated on open-domain datasets like news articles, Wikipedia, etc. As a result, they lack robust capabilities to handle enterprise-specific vocabulary and named entities. For instance, standard named entity recognition models fail to identify key industry terms, customer ids, product codes, etc. This leads to poor extraction of salient details from enterprise documents.

Likewise, out-of-the-box sentence embeddings also struggle to capture domain-specific semantics and relationships between entities in enterprise corpora. This affects identifying central information and generating cohesive summaries.

Therefore, there is a need for developing customized summarization techniques for enterprise content containing unique vocabulary and details not found in generic training data. Accurate extraction and abstraction of key details from vast enterprise documents can help knowledge management and onboarding.

* + 1. **Related Works**

Prior research on text summarization has focused extensively on open-domain datasets:

1. Standard news and Wikipedia datasets have been studied widely using methods like TF-IDF, TextRank, Seq2Seq, etc.
2. Neural summarization techniques using CNNs, LSTMs, and attention have been applied to these public datasets.
3. Evaluation involves ROUGE, METEOR, etc. on DUC benchmarks containing news, articles, etc.
4. However, these methods fail to generalize to enterprise domains with unique vocabulary and semantics.

Limited past work has explored summarization on specific domains like scientific papers, reviews, or electronic health records containing some unique entities. However, the datasets are still orders of magnitude simpler than enterprise corpora.

Our work aims to address the lack of research on adapting summarization systems to complex enterprise documents with company-specific details at scale. We customize techniques for robust entity extraction and abstractive summarization tailored to enterprise data.

* 1. **Aims and Objectives**

The overarching aim seems to be developing and evaluating natural language processing (NLP) techniques and models for various tasks like sentiment analysis, named entity recognition, and part-of-speech tagging across different languages.

Specific objectives include:

1. Performing cross-lingual sentiment analysis by machine translating datasets.
2. Exploring deep learning methods like LSTMs for detecting implicit aspects in sentiment analysis.
3. Developing and evaluating entity recognition pipelines using libraries like spaCy.
4. Creating open-source POS taggers and entity recognizers for languages like Greek using spaCy.
5. Comparing the performance of different NLP libraries/toolkits like spaCy, Stanford NLP, and NLTK for named entity recognition.
   1. **Research Questions**

* How effective are machine translation and cross-lingual techniques for sentiment analysis tasks?
* Can deep learning models like LSTMs improve the detection of implicit aspects in sentiment analysis?
* What are optimal pipelines and configurations for named entity recognition using spaCy?
* How does the performance of spaCy's models compare to other NLP libraries for tasks like NER?
* Can state-of-the-art results be achieved for POS tagging and NER in low-resource languages like Greek using spaCy?
  1. **Scope of the Study**

The primary scope of this study is to develop advanced natural language processing techniques and models specifically tailored for enterprise text summarization using the SpaCy library. The key subject to be investigated is applying and extending SpaCy's core NLP capabilities like tokenization, POS tagging, NER, embeddings, etc. to accurately process complex enterprise documents containing domain-specific terminology, entities, and semantics.

The theoretical constructs underpinning the study involve:

1. Customizing SpaCy's neural architectures and embedding layers to incorporate domain knowledge and vocabularies from enterprise data.
2. Extending named entity recognition models to identify and classify proprietary entities like product codes, customer IDs, etc.
3. Fine-tuning sentence/document representation models on enterprise corpora to capture domain relationships and semantics.
4. Designing robust summarization architectures tailored for the long, intricate structures of enterprise documents.
5. Exploring ways to integrate domain-specific rules and knowledge bases to enhance extractive and abstractive summary quality.

The study will be conducted by first analyzing the unique challenges of enterprise document summarization compared to open-domain settings. Relevant enterprise datasets across domains like finance, consulting, manufacturing, etc. will be collected.

SpaCy's core models will be systematically customized, extended, and fine-tuned on this enterprise data using techniques like transfer learning, multi-task learning, and domain adaptation. The performance of these customized models will be rigorously evaluated on benchmark summarization tasks against baseline open-domain systems.

The empirical results will provide insights into the effectiveness of the developed techniques in improving salient information extraction and generating coherent, accurate summaries of enterprise documents. Qualitative analyses of samples will also assess the models' robustness to domain vocabulary and semantics.

Ultimately, the outcome of this study will be customized SpaCy pipelines optimized for the mission-critical task of enterprise text summarization, unlocking knowledge locked in proprietary documents.

* 1. **Significance of the Study**

This study holds significant importance due to the critical need for accurate text summarization capabilities tailored for enterprise/proprietary content. Key significances include:

1. Unlocking knowledge assets: Automatically extracting salient information from vast enterprise documents like reports, memos, etc. can greatly benefit knowledge management and access to proprietary data.
2. Enabling enterprise language models: Developing domain-aware NLP models that understand industry-specific terminology, products, and entities is crucial for enterprise AI applications.
3. Improving employee productivity: Succinct summaries of lengthy technical/internal documents can drastically reduce manual effort for employees, aiding onboarding and decision-making.
4. Advancing transfer learning for NLP: Systematic techniques to customize pre-trained language models for domain-specific use cases can drive NLP research.
5. Benchmarking enterprise text summarization: Lack of standardized enterprise datasets and benchmarks limits progress - this study provides an evaluation framework.
6. **Literature Review**
   1. **Introduction**

Extractive summarization methods select the most important sentences/passages from the input text to form the summary. This relies on identifying salient content via techniques like scoring sentences based on cue words, topic signatures, centrality, etc. SpaCy's NLP components for tokenization, POS tagging, NER, parsing, etc. can provide rich linguistic features to drive these extractive models.

Abstractive methods go beyond extracting snippets - they aim to generate entirely new sentences that concisely convey the key information from the input text. This involves deeper language understanding and generation capabilities like semantic representations, abstractive reasoning, and surface realization. SpaCy's neural architectures, language models, and generation components are relevant for abstractive summarization.

Discuss the relative strengths, limitations, and use cases of extractive vs abstractive approaches. Also covers recent trends towards hybrid models that combine extraction and abstraction.

* + 1. **History of Text Summarization**

The history of text summarization dates back several decades, with significant advancements primarily driven by the increasing availability of digital text and the evolution of natural language processing (NLP) technologies.

1. **Early Beginnings (1950s-1960s):**
   * The concept of automatic text summarization emerged in the 1950s with the advent of digital computers.
   * Hans Peter Luhn, a researcher at IBM, was one of the pioneers in this field. In 1958, he introduced a method for generating abstracts from scientific and technical literature using word frequency techniques.
   * Early approaches were primarily extractive, focusing on selecting and extracting salient sentences from the text.
2. **Development of Statistical Methods (1970s-1980s):**
   * The 1970s and 1980s saw the development of more sophisticated statistical methods for text summarization.
   * Researchers began exploring features such as sentence position, term frequency, and cue words to identify important sentences.
   * The use of statistical models and heuristic rules became prevalent during this period.
3. **Introduction of Machine Learning (1990s-2000s):**
   * The 1990s marked the introduction of machine learning techniques in text summarization.
   * Machine learning models, such as Naive Bayes and decision trees, were employed to learn from examples and improve the quality of summaries.
   * The advent of the internet and the explosion of online content provided vast amounts of data for training summarization systems.
4. **Advancements in NLP and Deep Learning (2010s-Present):**
   * The past decade has seen significant advancements in NLP, particularly with the rise of deep learning.
   * Neural networks, especially recurrent neural networks (RNNs) and later transformer models, have revolutionized text summarization.
   * Abstractive summarization, which generates new sentences rather than extracting existing ones, has become more feasible with the advent of models like the Transformer, BERT, and GPT.
   * Techniques such as reinforcement learning and transfer learning have further enhanced the capabilities of summarization systems.
5. **Current Trends and Future Directions:**
   * Modern text summarization systems leverage pre-trained language models like GPT-3, T5, and BART, which have demonstrated remarkable performance in generating coherent and contextually relevant summaries.
   * There is a growing interest in domain-specific summarization, where models are fine-tuned on specific types of content, such as legal documents, scientific papers, or news articles.
   * Researchers are also focusing on improving the factual consistency and interpretability of summaries, ensuring they are not only concise but also accurate and trustworthy.

In conclusion, the field of text summarization has evolved significantly from simple frequency-based methods to sophisticated neural models capable of generating human-like summaries. As NLP technologies continue to advance, we can expect further improvements in the quality and applicability of text summarization systems.

* + 1. **Applications of Text Summarization**

Text summarization, the process of creating concise and coherent summaries from longer texts, has a wide array of applications across different domains. Here are some of the key applications:

1. **News Aggregation and Media**:
   * Summarization is widely used to generate short summaries of news articles, enabling readers to quickly grasp the main points without reading the entire content. This is crucial for news apps and websites where users prefer brief overviews.
2. **Academic Research**:
   * Researchers utilize text summarization to generate abstracts or summaries of academic papers. This helps in quickly identifying relevant studies and understanding the gist of large volumes of research papers.
3. **Legal Documents**:
   * Legal professionals use summarization tools to extract key points from lengthy legal documents, contracts, and case files, which streamlines the review process and saves time.
4. **Business Intelligence**:
   * Summarization is used to distill information from financial reports, market research, and internal documents. It aids in making informed decisions by highlighting critical information.
5. **Customer Service**:
   * Automated summarization helps in generating brief responses to customer queries by summarizing customer interactions, support tickets, and feedback forms, enhancing the efficiency of customer support teams.
6. **Healthcare**:
   * Summarization tools help in creating concise summaries of patient records, medical literature, and clinical trial reports. This aids healthcare professionals in making quick and informed decisions.
7. **Social Media Monitoring**:
   * Businesses and marketers use summarization to analyze and summarize social media content and trends. This helps in understanding customer sentiments and emerging trends without going through vast amounts of data.
8. **Educational Tools**:
   * Summarization is employed in educational platforms to provide brief overviews of textbooks and course materials, assisting students in grasping key concepts quickly.
9. **E-commerce**:
   * Product summarization helps in creating concise product descriptions from user reviews and detailed product information. This enhances the shopping experience by providing potential buyers with essential details.
10. **Document Management Systems**:
    * Organizations use summarization to manage and organize large volumes of documents by generating summaries that make it easier to retrieve and categorize information.

By leveraging advanced natural language processing techniques, text summarization tools are able to handle the nuances of different domains, ensuring that the summaries produced are both accurate and useful.

* + 1. **Recommendations of Text Summarization**

Text summarization is a rapidly evolving field with significant potential for enterprise applications. Here are several recommendations to enhance the effectiveness of text summarization techniques, with a focus on leveraging SpaCy for Named Entity Recognition (NER) and other summarization methods:

* + - 1. **Leverage Domain-Specific Knowledge**:

Integrate domain-specific ontologies and knowledge bases to enhance the accuracy and relevance of summaries. This can help disambiguate entities and improve the coherence of the generated summaries.

Use SpaCy’s capabilities to incorporate structured knowledge, which can aid in resolving co-references and understanding the context of domain-specific terms.

* + - 1. **Utilize Hierarchical Summarization Techniques**:

For long enterprise documents, implement hierarchical summarization methods. This involves summarizing individual sections or subsections first and then aggregating these summaries to form a comprehensive document summary.

SpaCy’s discourse parsing and hierarchical segmentation tools can be particularly useful for this approach.

* + - 1. **Enhance Salience Estimation**:

Develop robust salience estimation models to identify the most critical information within a document. Techniques such as graph-based methods, topic modeling, and clustering can be employed using SpaCy’s embeddings and semantic similarity features.

Integrate lexical cues and discourse features to improve salience scoring, ensuring that the most relevant content is prioritized in the summaries.

* + - 1. **Incorporate Multi-Task Learning**:

Utilize multi-task learning models that combine summarization with other related NLP tasks, such as question answering, translation, and NER. This can lead to better representations and improved performance across tasks.

SpaCy supports multi-task learning, enabling the development of models that leverage shared representations for summarization and other enterprise-specific tasks.

* + - 1. **Adapt Pre-trained Models to Specific Domains**:

Fine-tune pre-trained language models on domain-specific data to improve their performance on enterprise documents. Continued pretraining on enterprise data and fine-tuning on target domain texts can significantly enhance summarization quality.

Use SpaCy’s customizable language models to integrate domain-specific vocabularies and entities effectively.

* + - 1. **Improve Named Entity Recognition (NER)**:

Enhance NER capabilities to accurately identify and classify proprietary entities within enterprise documents. Techniques such as few-shot learning, pattern matching, and dictionary-based methods can be employed to recognize new entity types.

SpaCy’s NER tools can be fine-tuned to detect and link domain-specific entities, ensuring that critical information is accurately captured in the summaries.

* + - 1. **Implement Evaluation Metrics and Pipelines**:

Develop comprehensive evaluation pipelines using metrics such as ROUGE scores, factual consistency, and coherence/grammaticality to assess the quality of summaries.

SpaCy offers tools for semantic similarity scoring, factuality classification, and grammar checking, which can be integrated into evaluation frameworks.

* + - 1. **Promote SpaCy for NER and Summarization**:

SpaCy provides robust support for both extractive and abstractive summarization methods. Its powerful NER, POS tagging and dependency parsing capabilities make it well-suited for extractive techniques, while its neural architecture components are relevant for abstractive summarization.

By leveraging SpaCy’s advanced NLP features, enterprises can develop more accurate and context-aware summarization systems tailored to their specific needs.

In conclusion, adopting these recommendations can significantly improve the effectiveness of text summarization in enterprise settings. Utilizing SpaCy’s extensive NLP capabilities will further enhance the accuracy, coherence, and relevance of the generated summaries, providing valuable insights and facilitating better decision-making.

* 1. **Text Summarization Fundamentals**
     1. **Extractive vs Abstractive Methods**

SpaCy supports both extractive and abstractive text summarization approaches. Extractive methods identify and extract the most salient sentences from the input to form the summary. SpaCy's powerful NER, POS tagging, and dependency parsing capabilities can aid extractive techniques. Abstractive methods generate entirely new sentences for the summary, for which SpaCy's language modeling and neural architecture components are relevant.

* + 1. **Evaluation Metrics**

Summarizing the common quantitative metrics used to evaluate summarization systems:

* ROUGE scores: Variants like ROUGE-N, and ROUGE-L measure n-gram overlap between system and reference summaries
* Factual consistency: Ensuring summaries do not hallucinate incorrect details
* Coherence/Grammaticality: Scoring linguistic quality and fluency of summaries

Also discuss techniques in SpaCy that can aid in constructing evaluation pipelines, like semantic similarity scorers, factuality classifiers, grammar/style checkers, etc.

* + 1. **Key Challenges**

Some key challenges in text summarization and how SpaCy's capabilities can help:

* Long document summarization: Using SpaCy for hierarchical summarization, discourse analysis
* Domain transfer: Fine-tuning SpaCy models on target domain data
* Balancing salience and non-redundancy: SpaCy representation learning, MMR techniques
* Preserving key entities: Leveraging SpaCy's NER, entity linking
* Handling noisy data: SpaCy's data cleaning, denoising tools
  1. **Transfer Learning for Text Summarization**
     1. **Domain Adaptation Techniques**

Discuss methods to adapt SpaCy's pre-trained language models and NLP components to new domains, which is critical for enterprise settings:

* Continued pretraining on target domain: Further pretraining SpaCy language models on enterprise data
* Fine-tuning: Techniques to fine-tune SpaCy's summarization models on domain data
* Domain-aware pretraining: Incorporating domain entities/relations during SpaCy pretraining
* Data augmentation: Using SpaCy to generate synthetic enterprise data
  + 1. **Multi-Task Learning**

SpaCy supports multi-task learning models that leverage shared representations across related NLP tasks like summarization, QA, translation, etc.

Analyze approaches that combine summarization with auxiliary tasks relevant to enterprise content understanding:

* Multi-task with entity typing, and relation extraction to better capture enterprise knowledge
* Combining summarization and document-grounded question-answering
* Using summarization as an auxiliary task to improve domain language models
  1. **Customizing Language Models**
     1. **Vocabulary Integration Approaches**

Techniques to extend SpaCy's models to handle specialized vocabularies in enterprise domains:

* Subword tokenization to model infrequent technical terms
* Incremental vocabulary expansion with domain terms
* Character-level/hybrid models to compose unseen terms
* Few-shot learning methods to quickly integrate new vocabularies
  + 1. **Entity Recognition Extensions**

Enhancing SpaCy's named entity recognition to identify and classify proprietary entities:

* Few-shot learning approaches to recognize new entity types like product codes
* Pattern matching, dictionary-based techniques to detect domain entities
* Propagating entity labels from NER to language models for consistency
* Multi-task learning with related tasks like entity typing, relation extraction
  1. **Incorporating Domain Knowledge**
     1. **Knowledge Bases for NLP**

Discuss how existing enterprise knowledge bases and ontologies can be leveraged to improve summarization of domain-specific content:

* Using knowledge bases to disambiguate entities, resolve co-references across documents
* Injecting domain constraints from ontologies into summarization models
* Techniques to integrate structured knowledge with SpaCy's statistical representations
* Methods to automatically extend/learn from enterprise knowledge resources
  + 1. **Rule Injection Methods**

Explore different approaches to inject symbolic rules, heuristics, and constraints derived from domain expertise into SpaCy's neural summarization models:

* Rule-based decoding constraints to control summary content/factuality
* Posterior regularization to imbue rules into neural model behavior
* Combining rule-based and neural components in hybrid architectures
* Rationale techniques to analyze and refine injected rules

Discuss trade-offs between fully data-driven and more hybrid knowledge-aware approaches.

* 1. **Summarization for Long Documents**
     1. **Hierarchical Models**

Long enterprise documents like reports, manuals, etc. have an inherent hierarchical structure. Discuss hierarchical summarization techniques using SpaCy:

* Models that first summarize subsections before composing a higher-level summary
* Hierarchical segmentation/parsing of documents with SpaCy's discourse parsing
* Coarse-to-fine approaches to progressively summarize different granularities
* Leveraging existing document structure and section metadata
  + 1. **Salience Estimation**

Identifying the most salient information across a long document is crucial for summarization. Cover different salience estimation techniques with SpaCy:

* Graph-based methods using SpaCy's dependency parsing and semantic similarity
* Topic modeling and clustering approaches with SpaCy's embeddings/topic models
* Integrating lexical cues, and discourse features from SpaCy for salience scoring
* Few/zero-shot techniques to quickly adapt salience scorers to new domains

Also, explore summary-worthy section selection and content selection methods for long documents.

1. **Research Methodology**
   1. **Data Selection**
      1. **Purpose**

The data selection process aims to identify and gather a diverse set of documents relevant to the enterprise-specific text summarization task.

* + 1. **Approach**

Utilize a combination of manual and automated methods to curate a comprehensive dataset, including documents from various sources such as reports, articles, and internal documents.

* + 1. **Criteria**

Select documents based on relevance to the enterprise domain, diversity in content, and representativeness of the information landscape.

* 1. **Pre-processing**

Data preprocessing is essential to clean and prepare the raw text data for effective text summarization using SpaCy.

* + 1. **Tokenization**

Segment the text into tokens (words, punctuation, etc.) using SpaCy's tokenization capabilities. We employed SpaCy's tokenizer with custom rules to handle code snippets, technical terms, and domain-specific abbreviations correctly.

* + 1. **Selective Stop-word Removal**

Eliminate common stop words that do not add significant meaning to the text.

* + 1. **Lemmatization and Compound Word Handling**

We utilized SpaCy's lemmatizer and compound word handling capabilities to ensure an accurate representation of technical terms and programming constructs.

* + 1. **Part-of-Speech Tagging and Named Entity Recognition**

We leveraged SpaCy's part-of-speech tagging and named entity recognition components to identify and preserve domain-specific entities, such as product names, software components, and technical jargon.

* + 1. **Code Extraction and Formatting**

We implemented custom rules to extract and format code snippets from the documents, preserving their structure and syntax for subsequent analysis.

* 1. **Data Engineering**

To represent the diverse and highly technical nature of the data, we utilized a combination of techniques:

* + 1. **Word Embeddings**

We trained domain-specific word embeddings on our corpus of technical documents using SpaCy's word vector functionality. These embeddings capture the semantic and syntactic relationships between enterprise-specific terms and programming constructs.

* + 1. **Feature Engineering**

In addition to word embeddings, we extracted a range of features from the pre-processed data, including part-of-speech tags, named entities, code snippets, and contextual information (e.g., document metadata, section headings).

* + 1. **Data Augmentation**

To address potential data scarcity for specific domains or technical areas, we employed data augmentation techniques, such as synonym replacement and back-translation, using domain-specific dictionaries and resources.

These techniques enabled us to create rich and informative representations of enterprise-specific data, which can be effectively utilized by downstream machine learning models for information extraction and abstraction tasks.

* 1. **Interactive Visual Analytics**

Interactive Visual Analytics plays a crucial role in exploring and analyzing complex enterprise data for text summarization. Here is a detailed breakdown of the subtopics related to Interactive Visual Analytics

* + 1. **Semantic Similarity Visualization**

We leveraged dimensionality reduction techniques, such as t-SNE, to project the high-dimensional word embeddings into a 2D or 3D space. This visualization enables users to explore the semantic relationships between technical terms, programming constructs, and domain-specific concepts.

* + 1. **Topic Modelling and Clustering**

We employed topic modeling algorithms, such as Latent Dirichlet Allocation (LDA), to automatically identify and visualize topics and clusters within the corpus. This feature allows users to navigate the data based on high-level concepts and explore related technical information.

* + 1. **Code Exploration**

We developed interactive code visualization tools that enable users to explore code snippets within their context, navigate dependencies, and understand the relationships between different components and modules.

* **Functionality:** Developing interactive tools for visualizing code snippets within their context, navigating dependencies, and understanding relationships between different components.
* **Implementation:** Creating visual representations that allow users to explore and comprehend code structures and interconnections effectively.
* **Significance:** Enhances the understanding of technical details within the enterprise documents, aiding in knowledge extraction and analysis.
  + 1. **Collaborative Annotation**

Our platform supports collaborative annotation capabilities, allowing SMEs and technical staff to provide feedback, corrections, and additional context to the extracted information. These annotations are then incorporated into the system's knowledge base, improving its accuracy and relevance over time.

These interactive visual analytics capabilities empower users to gain insights into enterprise-specific data, identify knowledge gaps, and provide valuable feedback to refine the information extraction and abstraction process.

* 1. **Interpretation and Evaluation**

To evaluate the performance of our information extraction and abstraction system, we employed a combination of automatic metrics and human evaluation:

* + 1. **Precision, Recall, and F1-Score**

We calculated these standard information extraction metrics by comparing the system's output against a manually annotated ground truth corpus. This evaluation focused on the system's ability to accurately identify and extract relevant technical information, such as code snippets, domain-specific terms, and programming constructs.

* + 1. **Coherence and Readability Evaluation**

We conducted a human evaluation study, where SMEs and technical staff assessed the coherence, readability, and overall quality of the extracted and abstracted information. This evaluation ensured that the system's output was comprehensible and useful for knowledge transfer and onboarding purposes.

* + 1. **Knowledge Retention Assessment**

Knowledge Retention Assessment: We designed assessments to measure the effectiveness of the system in facilitating knowledge retention among new joiners and technical staff. These assessments evaluated the participants' understanding and retention of enterprise-specific concepts, best practices, and technical information after interacting with the system's output.

* + 1. **Error Analysis and Feedback Integration**

We performed a detailed error analysis by manually examining the system's output and categorizing the types of errors encountered. This analysis provided insights into the system's limitations and guided the development of targeted improvements. Additionally, we integrated user feedback and annotations from the collaborative annotation platform to continuously refine and enhance the system's performance.

1. **ANALYSIS**
   1. **Introduction**

In this chapter, the research provides a detailed analysis of the data and methodologies applied throughout the research. The analysis includes data preprocessing, exploratory data analysis (EDA), topic modeling, and various text summarization techniques. The aim is to offer insights into the data characteristics, the steps taken to clean and prepare the data, and the methods used to extract meaningful information from the text corpus.

* 1. **Dataset Description:**

The dataset used in this study consists of multiple documents from enterprise-specific sources. These documents include a variety of text types, such as technical reports, project documentation, and internal communication records. The dataset was pre-processed to extract relevant information, remove noise, and ensure consistency across different data sources.

* + 1. **Library Imports**

**Core Libraries**:

numpy, pandas, spacy, torch, sklearn, gensim, networkx, and transformers.

**Specific Functionality**:

Tokenization, stop words, TF-IDF vectorization, KMeans clustering, and LDA topic modeling.

**Loading SpaCy Model**:

The code loads the large English language model from SpaCy (en\_core\_web\_lg), which includes vectors, tokenization, and NER capabilities.

**Loading Data**:

The text corpus is read from a CSV file into a pandas DataFrame.

* 1. **Preprocessing and Annotation**
     1. **Custom Stop Words and Tokenizer**:

Additional stop words are defined, and a custom tokenizer from SpaCy is initialized.

* + 1. **Annotation Function**:

The preprocess\_and\_annotate() function processes each document by tokenizing, removing stop words, lemmatizing, POS tagging, extracting named entities, and computing word embeddings.

Preprocessed tokens are aggregated for each document, and token lists are compiled for further processing.

* + 1. **Storing Annotations**:

The annotated corpus is stored with tokens, POS tags, entities, and word embeddings.

* 1. **Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is a crucial step in understanding the underlying patterns and structures within the dataset. EDA helps in identifying significant features, outliers, and potential issues that need to be addressed before applying more sophisticated analysis techniques.

* + 1. **Topic Modeling with LDA:**

The code creates a dictionary and corpus using Gensim's corpora.Dictionary and doc2bow methods.

An LDA model is trained on the corpus to identify topics.

* + 1. **Abstractive Summarization with T5**:

**Model Initialization:**

The T5 model (t5-large) and tokenizer are loaded.

**Input Formatting:**

Text is formatted and tokenized for input into the T5 model.

**Summary Generation:**

Summaries are generated using beam search with early stopping to ensure high-quality output.

* + 1. **TextRank-based Text Summarization:**

**Graph Construction:**

(Implied) The code builds a graph of sentences for TextRank computation.

**Summary Extraction:**

Extracts top sentences based on PageRank scores

* + 1. **Hierarchical Summarization:**

**Recursive Summarization Function**:

The summarize\_section function recursively generates summaries by considering the syntactic structure.

**Summary Generation:**

The generate\_hierarchical\_summary function applies the recursive summarization to the document.

* 1. **Analysis and Recommendation**
     1. **Preprocessing and Annotation**:

The preprocessing function effectively integrates SpaCy's NER and POS tagging capabilities. Consider adding more detailed annotations, such as dependency parsing results, which can be useful for hierarchical summarization.

* + 1. **Topic Modeling:**

Using LDA for topic modeling is a robust choice for understanding document themes. Ensure the number of topics (num\_topics=5) aligns with the document's complexity. Experiment with different topic numbers to optimize results.

* + 1. **Abstractive Summarization:**

The T5 model is well-suited for abstractive summarization. Ensure the model\_max\_length and max\_length parameters are set appropriately for your text's length. You may need to preprocess the text to handle very long documents that exceed these lengths.

* + 1. **Extractive Summarization with TextRank:**

TextRank is a strong method for extractive summarization. Ensure the graph construction captures sentence similarities accurately. If not implemented, use TF-IDF vectors or cosine similarity for graph edge weights.

* + 1. **Hierarchical Summarization:**

The recursive approach for hierarchical summarization is innovative. Ensure the depth (max\_depth) parameter is tuned to balance summary conciseness and completeness.

* + 1. **Custom Stop Words and Tokenizer:**

Custom stop words improve preprocessing accuracy. Ensure the list comprehensively covers domain-specific stop words that may not be included in SpaCy's default set.

* 1. **Recommendations**
     1. **Evaluation Metrics:**

Incorporate evaluation metrics like ROUGE, BLEU, and METEOR to quantitatively assess summary quality.

* + 1. **Model Fine-Tuning**:

Consider fine-tuning the T5 model on domain-specific data for better performance.

* + 1. **Error Handling**:

Add error handling for cases where documents may not be processed correctly due to length or unexpected characters.

* + 1. **Scalability:**

Ensure the pipeline can handle large datasets efficiently, possibly by parallelizing parts of the code or optimizing data handling.

* + 1. **SpaCy Enhancements**:

Promote SpaCy's capabilities by demonstrating its flexibility in integrating with different summarization techniques, and fine-tuning its components for improved performance in specific tasks.

By following these recommendations and refining the code, you can build a robust, versatile text summarization system that leverages the strengths of both extractive and abstractive methods.

1. **Results and Discussions**

This section provides a comprehensive analysis of the results obtained from the research, interprets the visualizations, evaluates the sampling methods and their results, tests the validation dataset, and summarizes the findings.

* 1. **Introduction**

The Results and Discussions chapter is a critical part of the research, where the data collected is analyzed and interpreted. This section aims to present the findings in a structured manner, highlighting key insights and their implications. It begins with an overview of the data analysis process, followed by detailed interpretations of visualizations, evaluations of sampling methods, and validation testing, and concludes with a summary of the main findings.

* 1. **Interpretation of Visualizations**

Visualizations play a crucial role in understanding complex data sets. In this subsection, we will interpret the various visual representations of our data to uncover patterns, trends, and anomalies.

* + 1. **Graphs and Charts**:

These visual tools help in comparing different data sets and identifying trends over time. For instance, line graphs can show the progression of certain metrics, while bar charts can compare quantities across different categories.

* + 1. **Heatmaps**: Useful for identifying areas of high and low intensity within a data set, heatmaps can reveal correlations and cluster formations.
    2. **Scatter** **Plots**: These are employed to identify relationships between two variables. Clusters and outliers can be easily spotted, providing insights into data distribution.
    3. **Histograms**: By showing the frequency distribution of data, histograms help in understanding the spread and central tendency of the dataset.

Each visualization is analyzed to extract meaningful insights that contribute to our understanding of the data and support the research hypotheses.

* 1. **Evaluation of Sampling Methods and Results**

Sampling is a pivotal step in the research process, ensuring that the sample accurately represents the population. This section evaluates the sampling methods used and discusses their effectiveness.

* + 1. **Random Sampling**:

This method ensures that every member of the population has an equal chance of being selected. The results from random sampling are assessed for their representativeness and reliability.

* + 1. **Stratified Sampling**:

By dividing the population into subgroups and sampling each subgroup, this method aims to ensure that all relevant segments are adequately represented. The evaluation focuses on the precision and accuracy of the results.

* + 1. **Cluster Sampling**:

Often used in large populations, cluster sampling involves dividing the population into clusters and randomly selecting clusters for study. The effectiveness of this method is evaluated based on the homogeneity within clusters and heterogeneity between clusters.

The results from these sampling methods are compared and contrasted to determine the most effective approach for the research objectives.

* 1. **Testing of Validation Dataset**

Validation is essential to ensure that the model or analysis is generalizable to new data. This subsection discusses the process and outcomes of testing the validation dataset.

* + 1. **Data Splitting**: The original dataset is split into training and validation sets. The performance of the model on the validation set is crucial for assessing its accuracy and robustness.
    2. **Model Performance Metrics**: Metrics such as accuracy, precision, recall, and F1-score are used to evaluate the model’s performance. These metrics provide a comprehensive understanding of how well the model generalizes to unseen data.
    3. **Cross-Validation**: This technique involves dividing the data into several subsets and training the model multiple times, each time using a different subset as the validation set. Cross-validation helps in mitigating overfitting and provides a more reliable estimate of model performance.

The findings from the validation tests are discussed, highlighting any discrepancies and potential areas for improvement.

* 1. **Summary**

This subsection provides a concise summary of the key findings and discussions presented in this chapter.

* + 1. **Key Insights**:

Summarizes the main insights gained from the visualizations and their interpretations.

* + 1. **Sampling Methods**:

Reviews the effectiveness of different sampling methods and their impact on the results.

* + 1. **Validation Results**:

Highlights the performance of the model on the validation dataset and discusses its generalizability.

* + 1. **Implications**:

Discusses the implications of the findings for the research field and potential areas for future research.

The summary encapsulates the essence of the Results and Discussions chapter, providing a clear and comprehensive overview of the research findings and their significance.

1. **Conclusions and Recommendations**
   1. **Introduction**
   2. **Discussion and Conclusion**
   3. **Contribution to knowledge**
   4. **Future Recommendation**
2. **References**

**Appendix A: Research Proposal**

**Appendix B: Ethic Forms**

**Appendix C: Research Papers**

1. <https://ieeexplore.ieee.org/document/9498372> - Deep\_Learning\_WordNet\_and\_spaCy\_based\_Hybrid\_Method\_for\_Detection\_of\_Implicit\_Aspects\_for\_Sentiment\_Analysis
2. <https://ieeexplore.ieee.org/document/9197829> - Polarity\_Detection\_in\_a\_Cross-Lingual\_Sentiment\_Analysis\_using\_spaCy
3. <https://ieeexplore.ieee.org/document/10020570> - Twitter\_Accounts\_Suggestion\_Pipeline\_Technique\_SpaCy\_Entity\_Recognition
4. <https://ieeexplore.ieee.org/document/9404712> - Extractive\_Automatic\_Text\_Summarization\_using\_SpaCy\_in\_Python\_amp\_NLP
5. <https://ieeexplore.ieee.org/document/8931850> - A\_Replicable\_Comparison\_Study\_of\_NER\_Software\_StanfordNLP\_NLTK\_OpenNLP\_SpaCy\_Gate
6. <https://ieeexplore.ieee.org/document/10441408> - ROUGE\_Score\_Analysis\_and\_Performance\_Evaluation\_Between\_Google\_T5\_and\_SpaCy\_for\_YouTube\_News\_Video\_Summarization
7. <https://ieeexplore.ieee.org/document/9509650> - An\_Efficient\_method\_for\_Aspect\_Based\_Sentiment\_Analysis\_Using\_SpaCy\_and\_Vader
8. <https://ieeexplore.ieee.org/document/8909591> - Design\_and\_implementation\_of\_an\_open\_source\_Greek\_POS\_Tagger\_and\_Entity\_Recognizer\_using\_spaCy
9. <https://ieeexplore.ieee.org/document/9764935> - Morphological\_annotation\_of\_the\_Slovak\_language\_in\_the\_Spacy\_library\_with\_the\_pretraining