Does Anaphora Resolution Improve LLM Fine-Tuning for Summarisation?

Yi-Chun Lo[1] and **Ruslan Mitkov**[2]
[1]Lancaster University, United Kingdom
[2]University of Alicante, Spain
y.lo4@lancaster.ac.uk ruslan.mitkov@ua.es

Abstract

This study investigates whether adding anaphora resolution as a preprocessing step before fine-tuning the text summarisation application in Large Language Model (LLM) can improve the quality of summary output. We conducted two sets of training with the T5-base model and BART-large model using the SAMSum dataset. One used the original text and the other used the text processed by a simplified version of MARS (Mitkov's Anaphora Resolution System). The experiment revealed that when T5-base model was fine-tuned on the anaphora-resolved inputs, the ROUGE-1, ROUGE-2 and ROUGE-L metrics were improved from 45.8567, 22.0195 and 38.0433 to 48.0281, 24.4447 and 40.3584 respectively (Wilcoxon signed-rank test p-value less than 0.01 and paired t-test p-value less than 0.01). In contrast, BART-large model only had a slight improvement after fine-tuning under the same conditions, which was not statistically significant. Further analysis of the generated summaries confirmed that anaphora resolution was helpful in semantic alignment. In conclusion, this study demonstrates that adopting anaphora resolution as a preprocessing step for LLM fine-tuning is effective in enhancing the performance of summarisation in T5-base model. Although it did not reach statistical significance on BART-large, it still has practical value for small LLM or scenarios with limited computing resources.

1 Introduction

In recent years, the rapid development of Large Language Model (LLM) has greatly contributed to the advancement of various areas in Natural Language Processing (NLP). With the increasing ability of these models to understand and generate language, text summarisation is an important and widely used application that increasingly relies on LLM for processing. Whether it is news summarisation, meeting record organisation, or social media content compression, LLM has demonstrated a strong ability to generate summaries (Gusev, 2020; Pan et al., 2024; Blekanov et al., 2022).

To further improve the performance of LLM on specific tasks, fine-tuning is one of the most common strategies. By fine-tuning on the downstream task dataset, the model can better adapt to the target task and improve the quality of the output. However, the effect of fine-tuning depends not only on the model structure design and training arguments, but also on the characteristics of the input data. In this background, anaphora resolution is particularly important. It refers to the automatic identification of the antecedent to which an expression (such as a pronoun or a noun phrase) in a text refers, and is an essential part of language interpretation. As Mitkov (2002) pointed out, anaphora resolution is a vital task for computers to comprehend natural language. Nevertheless, most of the past studies have focused on the internal evaluation of the anaphora resolution itself or analysing its overall impact on specific applications. Mitkov et al. (2007, 2012) have also investigated the results of anaphora resolution and coreference resolution (not only backward-pointing references but includes all mentions referring to the same entity) in NLP applications, and indicated that anaphora resolution can bring some degree of performance improvement. However, there is no systematic study to explore whether anaphora resolution as a data preprocessing step can significantly improve the fine-tuning effect of LLM. Therefore, in this paper, we conduct experiments aiming at the following core question:

> Can anaphora resolution preprocessing improve LLM summarisation fine-tuning?

Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models associated with RANLP 2025, pages 59–66, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-102-8-007

2 Related Work

Before understanding anaphora resolution, it is crucial to clarify the basic concept of *anaphor*, which is a word or phrase that points back to a previous reference in a discourse, such as a personal pronoun (he, she, it) or a definite noun phrase. In contrast, antecedent is the previous entity referenced by anaphor, usually in noun phrase (NP). Take the sentence mentioned by Mitkov (2022) as an example:

> *The Queen* said the UK will succeed in its fight against the coronavirus pandemic, in a rallying message to the nation. *She* thanked people for following government rules to stay at home.

In this case, *She* is the anaphor and *the Queen* is the antecedent, which establishes semantic relationship in the discourse. Anaphora resolution is the process for identifying the antecedent of an anaphor. Among some early approaches, Lappin and Leass (1994) developed an algorithm based on syntactic structures and heuristic rules that effectively combines semantic and discourse information for anaphora resolution. Ge et al. (1998) introduced a statistical approach to the construction of anaphora resolution decision tree using a data-driven method.

Mitkov (1998); Mitkov et al. (2002) proposed a different approach to knowledge-poor anaphora resolution. This method was later evolved into MARS (Mitkov's Anaphora Resolution System), which is a fully automated system for anaphora resolution. MARS has the advantage of simplicity, fast operation, and the ability to achieve about 60% accuracy in technical manuals without relying on knowledge bases.

In addition to discourse-level preprocessing techniques, modern text summarisation applications rely heavily on pre-trained LLMs. Early on, the Sequence-to-Sequence (Seq2Seq) neural network summarisation model (Sutskever et al., 2014) was developed. This model applies an 'encoder-decoder' framework to encode the entire text before generating a summary. In simple terms, the entire paragraph is encoded into a vector. The decoder then uses this vector and the generated words to generate a summary word by word. However, the vector cannot accommodate long texts, and key information from the beginning can be easily missed. Therefore, many studies have incorporated 'attention' into the encoder-decoder architecture (Bahdanau, 2014; Rush et al., 2015; Luong et al., 2015). During encoding, the state of each position is output. For each generated word, the decoder calculates a set of attention weights to focus on the most relevant positions. Subsequent research has incorporated the Transformer (Vaswani et al., 2017), using multi-head self-attention to model the entire text. In the encoder, self-attention is used to enable each word in a text to look back at other words in the text and determine which to focus on at the moment. The decoder uses masked self-attention to focus only on the generated portion, and cross-attention to allow the model to consider the most relevant parts of the original text when outputting the summary.

Among them, T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2020) is a representative Transformer model. In addition to the architecture mentioned above, the core concept of T5 is span corruption. During pre-training, a continuous segment of text is first removed from the original source, prompting the model to reconstruct the omitted passage. This is like asking the model to understand the context and fill in the missing content with its own words, just like the ability to read and retell the text required for summarisation. The design is not only flexible, but also allows it to perform well on a variety of summary datasets (Zhang et al., 2020; Hasan et al., 2021; Guo et al., 2021).

In contrast, BART (Bidirectional and Auto-Regressive Transformers) (Lewis et al., 2019) is another representative Transformer model. Unlike T5, in addition to removing consecutive segments, BART also utilises a denoising autoencoder to scramble the input before requiring the model to recover it. This is done to train the model to have greater understanding and reconstruction capabilities. This destruction-reconstruction method also enables BART to perform well on summary tasks (Yu et al., 2020; Yadav et al., 2023).

However, most studies have focused on the optimisation of the model itself, and have rarely explored the need for semantic enhancement of the input data in the fine-tuning process. Therefore, this is exactly the problem that this study aims to investigate.

3 Data

The dataset used in this study is SAMSum Corpus (Gliwa et al., 2019), a manually annotated conversation summary dataset of simulated two-person real-time chats in everyday life. There are more than 16,000 conversations in this dataset, each containing multiple rounds of speech with corresponding concise summaries. The dialogues are written and annotated by linguists, with a clear semantic structure and consistent style. The dataset is widely used in summarisation research and is one of the most common standardised assessment corpora available.

SAMSum is particularly suitable for this study due to the following reasons. Firstly, the data are multi-round spoken dialogues with a large number of pronouns, which are very likely to be ambiguous, and this is exactly the context in which anaphora resolution can be useful. Secondly, the output summaries of SAMSum are all abstractive style, so the model needs to have a deep understanding of semantics and discourse coherence in order to produce high quality summaries. By comparing the effect of fine-tuning before and after anaphora resolution, the effect of discourse clarity on model learning can be effectively observed. Although other datasets such as MeetingBank (Hu et al., 2023) and CNN/DailyMail (Nallapati et al., 2016) were also considered, most of these datasets do not have the conversational interactivity of SAMSum and do not require as much to identify antecedents in summaries. Furthermore, these datasets are larger than SAMSum. Given limited computing resources, SAMSum may be the most cost-effective choice.

However, the dataset has some limitations. As the conversations are simulated, they may not be as natural as real social platform conversations, and the scenarios are relatively focused on everyday conversations, which lacks topic diversity. Nevertheless, SAMSum is highly representative in terms of data quality, annotation consistency and task relevance, and is a suitable test to assess whether LLM benefits from discourse-level preprocessing such as anaphora resolution.

4 Methodology

The methodology of this study is divided into two stages. Firstly, anaphora resolution is performed on the dialogue texts of the training set in the SAMSum dataset using a self-implemented simplified version of MARS, in which the anaphor are replaced by their inferred antecedents. Then, T5 and BART models are fine-tuned using the anaphora resolution and the unprocessed versions of the data. Finally, by comparing the performance of the models in generating summaries on the test set, we analyse whether introducing anaphora resolution in data preprocessing can effectively improve the performance of the summarisation. In other words, we start with LLM that has been pre-trained on a large-scale corpus. To help the model learn to output summaries based on inputs, we fine-tune it on the SAMSum dataset, aligning its generated distribution with the target summaries. After training, during inference and testing, the model employs an autoregressive approach, conditioning on previously generated tokens to generate the next token. This study aims to investigate whether performing anaphora resolution on the SAMSum dataset during the fine-tuning phase can improve the final summarisation performance of the model.

4.1 Anaphora Resolution with MARS

In this study, a simplified version of MARS (Mitkov's Anaphora Resolution System) is used, with the core logic continued from the framework of Mitkov et al. (2002), which is approximately the same as its five processing phases. First, the system applies the FDG Parser from Conexor (Tapanainen and Jarvinen, 1997) to perform part-of-speech (POS) tagging, lemmatisation, and dependency parsing on the input text to extract compound NPs for subsequent use. Then, in the second stage, the system identifies potential referential pronouns and filters out non-referential 'it' by the machine learning method developed by Evans (2001). In the third stage, for each identified referential pronoun, NPs are selected as antecedent candidates from the heading of the paragraph, the current sentence and the first two sentences. Further filtering is performed according to grammatical constraints, requiring gender and number agreement between candidates and pronouns, and excluding grammatically impossible combinations. The fourth stage applies a set of antecedent indicators to all qualified candidates, which contain a total of 14 preferential and impeding factors, and each candidate receives a set of scores based on these indicators to measure its likelihood of becoming an antecedent. Finally, in the fifth stage, the candidate with the highest total score is chosen as the antecedent of the anaphor. In case of a tie, the most recent highest-scoring candidate is chosen.

However, there are many differences in the implementation details. First, in the syntactic analysis stage, considering the open source and efficiency issues, spaCy (Honnibal et al., 2020) is used to replace the original FDG Parser to perform POS tagging, dependency parsing, and to count the frequency of occurrence for NPs. In the second stage of pleonastic it filtering, the machine learning classifier proposed by Evans (2001) is abandoned and part of the discrimination rule proposed by Paice and Husk (1987) is applied instead. For the third stage of candidate extraction, the gender agreement check is omitted because of the uncertainty in the correspondence between names and genders in the conversation dataset and the high risk of gender mismatch. During the fourth stage, the original 14 indicators other than boost pronoun are employed. However, collocation match only compares the lemma without creating a collocation database, and term preference replaces the original TF-IDF method with the highest-frequency occurring NPs. In addition, instead of implementing a Genetic Algorithm (GA) for automatic weight optimisation (OrĂsan et al., 2000), the system adopts a fixed score, which is expected to run in a more stable and lighter way.

4.2 Fine-Tuning Setup

This study utilises T5-base and BART-large. T5-base is a publicly available version of the intermediate pre-training model in the T5 architecture, which has about 220M parameters with a complete encoder-decoder structure. BART-large is a high-level pre-training model based on the BART architecture, including a 12-layer encoder and a 12-layer decoder, with a total of approximately 402M parameters. These models strike a balance between resource consumption and model performance. In addition, this model selection can also take into account the variations in the scale of two parameters and test the performance of models with different structures. The original version of the SAMSum dataset has been divided into training and testing sets, so this study directly follows its default partitioning for model training and testing without any additional adjustment. We have designed two sets of inputs. One is the original dialogue data and the other is the anaphora-resolved version by MARS. Each is used to fine-tune models with the same structure and settings, so that a fair comparison can be made as to whether anaphora resolution improves model summarisation.

For the training arguments, the batch size is set to 8, the learning rate is set to 0.0001, and the training is conducted with 3 epochs. In order to retain some of the pre-training knowledge and reduce the consumption of resources, the weights of the first three encoder layers in both T5-base and BART-large are frozen. The optimiser employs AdamW (Loshchilov and Hutter, 2017) with a linear scheduler, where the learning rate decreases as the training progresses. Moreover, ROUGE-1, ROUGE-2 and ROUGE-L are considered as the summary quality assessment metrics in the test set (Lin, 2004).

In order to verify the differences in summary quality between different input versions are not due to random fluctuations, this study conducts Wilcoxon signed-rank tests (Wilcoxon, 1992) and paired Student's t-tests on the ROUGE-1, ROUGE-2, and ROUGE-L metrics of each sample in the test set. All tests are one-tailed, with the alternative hypothesis that the anaphora-resolved result increases higher ROUGE metrics. Furthermore, the Holm–Bonferroni method (Holm, 1979) is used to correct the multiple comparison results of the three metrics, with the significance level set to 0.01.

To ensure the reproducibility of our experiments, we set the number of random seeds to 413, and use the L4 GPU of Google Colab for training.

5 Results

After anaphora resolution on the SAMSum dataset, 2,479 (91.679%) of the 2,704 target pronouns were replaced. Consistent with the original MARS, the antecedent candidates in this study were restricted to the current sentence and the two preceding sentences. Of these replaced pronouns, approximately 48.81% had their antecedents in the same sentence, 31.18% in the previous sentence, and 20.01% in the previous two sentences. On average, each dialogue contained 3.3 pronouns. Anaphora resolution only slightly altered the input length, increasing each dialogue (per sample) by an average of 1.3056 tokens and 35.6174 characters. Moreover, this section reports the performance of the four fine-tuned models on the SAMSum test set in turn. First, two sets of results are presented for T5-base (original vs. resolved), and then two sets of results for BART-large (original vs. resolved).

5.1 T5-base

Table 1 lists the ROUGE metrics of the T5-base model on original input and the anaphora-resolved input. From the results, it could be seen that with the integration of anaphora resolution, the model showed significant improvement in all three ROUGE metrics. The one-tailed Wilcoxon signed-rank test (W-test) and the paired Student's t-test (t-test) results including test statistics, raw p-values, and Holm–Bonferroni adjusted p-values are reported in Tables 2, 3 and 4. All three ROUGE scores had p-values close to zero, confirming that the performance improvement brought by anaphora resolution is highly significant.

Model	ROUGE-1	ROUGE-2	ROUGE-L
Raw	45.8567	22.0195	38.0433
Resolved	**48.0281**	**24.4447**	**40.3584**

Table 1: ROUGE comparison for T5-base

Test	ROUGE-1	ROUGE-2	ROUGE-L
W-test	154033.50	127586.00	151217.50
t-test	6.31	6.04	6.08

Table 2: Test statistics for Wilcoxon signed-rank test (W-test) and paired Student's t test (t-test) on T5-base ROUGE metrics

Test	ROUGE-1	ROUGE-2	ROUGE-L
W-test	0	0	0
t-test	0	0	0

Table 3: Raw p-values for Wilcoxon signed-rank test (W-test) and paired Student's t-test (t-test) on T5-base ROUGE metrics

A dialogue from the SAMSum test set further demonstrated the semantic contrast between the two models. The summaries generated from the original model were compared with those from the anaphora-resolved model, as well as the artificial reference summaries. In this dialogue, Igor expresses his workload and depression during the two weeks before leaving his job, and John gives advice and counselling. However, the summary generated by the original model only mentioned that Igor was overloaded with work and focused on the persuasion of John to 'stop thinking and start doing'. It completely ignored the frustration of Igor and the assessment of John that it was irresponsible to assign too much work during the notice period. In contrast, the model summary after anaphora resolution not only captured the 'demotivated' mood of Igor, but also correctly reflected the criticism of excessive work allocation by John. This allowed the generated content to take into account both emotions of Igor and opinions of John, and was closer to the dual narrative of the reference summary. The full dialogue and model outputs can be found in Appendix A.1.

Test	ROUGE-1	ROUGE-2	ROUGE-L
W-test	0	0	0
t-test	0	0	0

Table 4: Holm–Bonferroni corrected p-values for Wilcoxon signed-rank test (W-test) and paired Student's t test (t-test) on T5-base ROUGE metrics

5.2 BART-large

Table 5 lists the ROUGE metrics of the BART-large model on original input and the anaphora-resolved input. From the overall trend, BART-large had slightly increased in all three ROUGE metrics after anaphora resolution, indicating that the semantic consistency of the generated summary has improved. The one-tailed Wilcoxon signed-rank test (W-test) and the paired Student's t-test (t-test) results including test statistics, raw p-values, and Holm–Bonferroni adjusted p-values are reported in Tables 6, 7 and 8. However, the results indicate that these improvements are not statistically significant. The p-values of these three scores are all greater than 0.01.

Model	ROUGE-1	ROUGE-2	ROUGE-L
Raw	49.6463	26.5392	41.9366
Resolved	**50.0213**	**26.8944**	**42.1020**

Table 5: ROUGE comparison for BART-large

Test	ROUGE-1	ROUGE-2	ROUGE-L
W-test	109756.50	91295.00	111061.00
t-test	0.12	0.23	0.05

Table 6: Test statistics for Wilcoxon signed-rank test (W-test) and paired Student's t test (t-test) on BART-large ROUGE metrics

On the semantic level, the BART-large model also showed obvious differences on the same test examples in Section 5.1. The original model mentioned that John suggested Igor to do what he had to do. The model after anaphora resolution clearly conveyed the view of John that it was irresponsible to assign too much work during the notice period. The full dialogue and model outputs can be found in Appendix A.2.

Test	ROUGE-1	ROUGE-2	ROUGE-L
W-test	0.3909	0.3399	0.5790
t-test	0.4520	0.4106	0.4798

Table 7: Raw p-values for Wilcoxon signed-rank test (W-test) and paired Student's t test (t-test) on BART-large ROUGE metrics

Test	ROUGE-1	ROUGE-2	ROUGE-L
W-test	1	1	1
t-test	1	1	1

Table 8: Holm–Bonferroni corrected p-values for Wilcoxon signed-rank test (W-test) and paired Student's t test (t-test) on BART-large ROUGE metrics

6 Discussion

This study confirms that adding anaphora resolution before fine-tuning can significantly improve the summary quality of the T5-base model, reaching high significance in all ROUGE metrics. For BART-large, although there was a small gain, it did not pass the significance test, indicating that its marginal benefit on large models is relatively limited.

The actual summary examples also confirmed the above results. The summary of the T5-base model without anaphora resolution only focuses on the heavy workload and ignores the emotional clues. After anaphora resolution, it can fully present the frustrated state of Igor. Although BART-large can add details about the evaluation of John for over-allocation of work and irresponsibility after anaphora resolution, the overall summary quality does not change much.

We believe that this difference stems from three main factors. First, replacing ambiguous pronouns with explicit noun phrases can greatly reduce the ambiguity of the input and facilitates direct alignment of semantic roles. Second, strengthening the coherence of the text allows the model to learn the correspondence between characters and context more efficiently. For small models, this lightweight preprocessing can significantly reduce the noise during fine-tuning and improve learning effects. Third, the model does not have to remember or learn the antecedents corresponding to different pronouns during training, and perhaps self-attention can be aligned without having to span large distances. However, for models with larger capacity and deep context modeling capabilities, the benefits are relatively diminishing. Moreover, we speculate that pre-training method of destroying the input enables BART to strengthen its understanding of entity and paragraph coherence during the reconstruction process, so the marginal benefit of anaphora resolution is relatively small compared to T5.

7 Conclusions and Future Work

This study investigates whether preprocessing with anaphora resolution before LLM fine-tuning for summary application can improve the model performance. By fine-tuning the T5-base model and the BART-large model on the SAMSum dataset with the original text and the text processed by the simplified version of MARS. The results show that T5-base achieves highly significant gains in ROUGE-1, ROUGE-2, and ROUGE-L metrics after anaphora resolution, which fully demonstrates how anaphora resolution enhances the ability of the model to capture semantic coherence. BART-large, on the other hand, only shows a small and non-significant increase in each metric, indicating that its innate contextual understanding already covers most parsing relationships, and thus has limited marginal benefits.

This study is still limited to the SAMSum dataset and two models. The applicability of other corpora, languages, or larger-scale LLMs remains to be verified. In addition, the interaction between hyperparameters (such as learning rate, number of frozen layers) and the benefits of anaphora resolution also needs to be systematically explored. Future research can further expand to more models and datasets. For example, at the model level, experiments can be conducted using larger LLMs such as GPT-NeoX-20B (Black et al., 2022) or Llama 2 (Touvron et al., 2023). At the data level, different styles and topics of summary datasets such as MeetingBank (Hu et al., 2023) or CNN/DailyMail (Nallapati et al., 2016) can be considered. Furthermore, according to a comparative study by Mitkov and Ha (2024), the use of state-of-the-art anaphora resolution methods based on deep learning (such as DeBERTa-based token labelling) may further improve the accuracy, which in turn may lead to stronger summarisation performance.

Acknowledgments

This work has been partially supported by the CIDEXG/2023/12 project, funded by the Generalitat Valenciana.

References

Dzmitry Bahdanau. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

Ivan S Blekanov, Nikita Tarasov, and Svetlana S Bodrunova. 2022. Transformer-based abstractive summarization for reddit and twitter: single posts vs. comment pools in three languages. *Future Internet*, 14(3):69.

Richard Evans. 2001. Applying machine learning toward an automatic classification of it. *Literary and linguistic computing*, 16(1):45–58.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Sixth workshop on very large corpora*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916*.

Ilya Gusev. 2020. Dataset for automatic summarization of russian news. In *Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings 9*, pages 122–134. Springer.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. spacy: Industrial-strength natural language processing in python.

Yebowen Hu, Tim Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. Meetingbank: A benchmark dataset for meeting summarization. *arXiv preprint arXiv:2305.17529*.

Shalom Lappin and Herbert J Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Ruslan Mitkov. 2002. *Anaphora resolution*. Routledge.

Ruslan Mitkov. 2022. Anaphora resolution. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, second, substantially revised edition, pages 707–729. Oxford University Press.

Ruslan Mitkov, Richard Evans, and Constantin Orasan. 2002. A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3*, pages 168–186. Springer.

Ruslan Mitkov, Richard Evans, Constantin Orăsan, Iustin Dornescu, and Miguel Rios. 2012. Coreference resolution: To what extent does it help nlp applications? In *Text, Speech and Dialogue: 15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012. Proceedings 15*, pages 16–27. Springer.

Ruslan Mitkov, Richard Evans, Constantin Orăsan, Le An Ha, and Viktor Pekar. 2007. Anaphora resolution: To what extent does it help nlp applications? In *Anaphora: Analysis, Algorithms and Applications: 6th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007, Lagos, Portugal, March 29-30, 2007. Selected Papers 6*, pages 179–190. Springer.

Ruslan Mitkov and Le An Ha. 2024. Are rule-based approaches a thing of the past? the case of anaphora resolution. *Procesamiento del Lenguaje Natural*, 73(0):15–27.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Constantin OrĂsan, Richard Evans, and Ruslan Mitkov. 2000. Enhancing preference-based anaphora resolution with genetic algorithms. In *Natural Language Processing—NLP 2000: Second International Conference Patras, Greece, June 2–4, 2000 Proceedings 2*, pages 185–195. Springer.

Chris D Paice and Gareth D Husk. 1987. Towards the automatic recognition of anaphoric features in english text: the impersonal pronoun "it". *Computer Speech & Language*, 2(2):109–132.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, et al. 2024. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. *arXiv preprint arXiv:2403.12968*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Pasi Tapanainen and Timo Jarvinen. 1997. A non-projective dependency parser. In *Fifth Conference on Applied Natural Language Processing*, pages 64–71.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer.

Hemant Yadav, Nehal Patel, and Dishank Jani. 2023. Fine-tuning bart for abstractive reviews summarization. In *Computational Intelligence: Select Proceedings of InCITe 2022*, pages 375–385. Springer.

Tiezheng Yu, Dan Su, Wenliang Dai, and Pascale Fung. 2020. Dimsum@ laysumm 20: Bart-based approach for scientific document summarization. *arXiv preprint arXiv:2010.09252*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

A Example Dialogue and Model Outputs

Input Dialogue

Igor: Shit, I've got so much to do at work and I'm so demotivated.
John: It's pretty irresponsible to give that much work to someone on their notice period.
Igor: Yeah, exactly! Should I even care?
John: It's up to you, but you know what they say...
Igor: What do you mean?
John: Well, they say how you end things shows how you really are...
Igor: And not how you start, right?
John: Gotcha!
Igor: So what shall I do then?
John: It's only two weeks left, so grit your teeth and do what you have to do.
Igor: Easy to say, hard to perform.
John: Come on, stop thinking, start doing!
Igor: That's so typical of you! ;)

Reference Summary *Igor has a lot of work on his notice period and he feels demotivated. John thinks he should do what he has to do nevertheless.*

A.1 T5-base

Summary from Raw Model *Igor has a lot of work to do. John advises him to stop thinking and start doing.*

Summary from Anaphora-Resolved Model *Igor has a lot of work to do. He is demotivated. John thinks it's irresponsible to give that much work to someone on their notice period.*

A.2 BART-large

Summary from Raw Model *Igor has a lot of work to do at work. John advises him to do what he has to do.*

Summary from Anaphora-Resolved Model *Igor has a lot of work to do at work. John reckons it's irresponsible to give so much work to someone on their notice period.*