Does Anaphora Resolution Improve LLM Fine-Tuning for Summarization?

Yi-Chun Lo
Lancaster University
y.lo4@lancaster.ac.uk

Abstract

This study investigates whether adding anaphora resolution as a preprocessing step before fine-tuning the text summarisation task in Large Language Model (LLM) can improve the quality of summary output. We conducted two sets of training with the T5-base model using the SAMSum dataset. One uses the original text and the other uses the text processed by a simpified version of MARS (Mitkov's Anaphora Resolution System). The experiment reveals that the version with anaphora resolution has significant improvement in ROUGE-1, ROUGE-2 and ROUGE-L metrics. Further analysis of the generated summaries also confirms that anaphora resolution is helpful in avoiding subject mismatch and semantic confusion. 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 tasks.

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 task 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 will conduct experiments aiming at the following core question:

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

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 pan-

demic, 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 is divided into five stages. First, the system applies the FDG Parser from Conexor (Tapanainen and Jarvinen, 1997) to perform part-of-speech (POS) tagging, lemmatization, 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. 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 tasks rely heavily on pre-trained LLMs. Among them, T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2020) is a representative model. The core concept of T5 is to unify all NLP tasks into a text-to-text format, which is designed to make the model more easy for transfer learning and more suitable for task generation. T5 also adopts an encoder-decoder architecture, which allows the model to learn how to reconstruct and understand the language by masking part of the text in a sentence during the pre-training stage. This 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). 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 tasks 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. 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, the T5 model is fine-tuned using the anaphora resolution and the unprocessed versions of the data. Finally, by comparing the performance of the two 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 task.

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. 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, 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 that strikes a balance between resource consumption and model performance. 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 layers in the T5 encoder are frozen. The optimiser employs AdamW (Loshchilov and Hutter, 2017) with a linear scheduler, where the learning rate decreases as the training progresses. Finally, ROUGE-1, ROUGE-2 and ROUGE-L are considered as the summary quality assessment metrics in the test set (Lin, 2004). The briefly explanation of there metrics can be found in Appendix A.

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

5 Results

Table 1 compares the performance of two models with different input data on the SAMSum test set after training. From the results, it can be seen that

Model	ROUGE-1	ROUGE-2	ROUGE-L
Raw	27.37	9.45	23.98
Anaphora-Resolved	**41.34**	**18.24**	**34.67**

Table 1: ROUGE comparison

with the integration of anaphora resolution, the model shows significant improvement in all three ROUGE metrics, especially in ROUGE-2, where the improvement is almost two times.

A dialogue from the SAMSum test set further demonstrates the semantic contrast between the two models. The summaries generated from the

original model are 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 original model incorrectly confuses the subject with the context and mistakenly assumes that it is John who is working. This resulted in semantic inconsistencies. With the addition of anaphora resolution, the model successfully identifies Igor as a worker and explicitly mentions the remaining two weeks of information, which is more closely aligned with the reference summary. The full dialogue and model outputs can be found in Appendix B.

6 Findings

The results of this study indicate that anaphora resolution before LLM fine-tuning significantly improves the performance of the summary task. The model with anaphora resolution outperforms the original model on all ROUGE metrics. The most significant improvement is in ROUGE-2, which shows a substantial improvement in the ability of the model to capture the coherence and composition of semantic units. Moreover, from the actual summaries generated by the model in the test set, it can be seen that the unprocessed model does not correctly map the anaphor to its antecedents. In contrast, after anaphora resolution, the model clearly indicates the subject and remaining time information, which is closer to the core meaning expressed in the reference summaries.

These findings can be attributed to the fact that anaphora resolution improves semantic clarity and consistency of the input text. First, by replacing pronouns such as he, she, and it with their corresponding noun phrases, the model can directly align semantic roles without further contextual reasoning, reducing ambiguity and mis-alignment during training. Second, the replacement of anaphor with antecedent strengthens discourse coherence and enables the model to learn semantic associations more efficiently. Overall, anaphora resolution not only provides a clearer input corpus, but also reduces the noise in the LLM learning process, which brings substantial benefits to the quality of the summaries.

7 Conclusions and Future Work

This study investigates whether preprocessing with anaphora resolution before LLM fine-tuning for summary task can improve the model performance. By fine-tuning the T5 model on the SAMSum dataset with the original text and the text processed by the simplified version of MARS. We observed significant performance differences. The experiment shows that the fine-tuned model with anaphora resolution outperforms the unprocessed version in ROUGE-1, ROUGE-2, and ROUGE-L metrics. In particular, ROUGE-2 shows the most significant improvement, highlighting its benefit in capturing the coherence and composition of semantic units.

We believe that anaphora resolution reduces the reliance of the model on contextual reasoning. During the training process, it helps the model to learn the association between characters and contexts more efficiently, and enhances the overall quality of the summaries. However, this study still has some limitations. Firstly, the experiments are only validated on a summary dataset (SAMSum) with a medium-sized pre-trained model (T5-base), and their generalisability and stability have not yet been evaluated on other data or larger models. Secondly, this study also does not systematically explore the effects of hyperparameter settings such as learning rate or number of frozen layers on the benefits of anaphora resolution.

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 BART (Lewis et al., 2019), 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–art anaphora resolution methods based on deep learning (such as DeBERTa-based token labelling) may further improve the parsing accuracy, which in turn may lead to stronger summarisation performance. This is worth exploring systematically in the future.

References

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, and

1 others. 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*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and 1 others. 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.

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

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. *The Oxford handbook of computational linguistics*. 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. *Proces. del Leng. Natural*, 73:15–27.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, and 1 others. 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, and 1 others. 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.

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, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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 ROUGE Metrics

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of evaluation metrics for summarisation tasks proposed by Lin (2004). Its core concept is to evaluate the quality of generated summaries by comparing the degree of content overlap between machine generated and human referenced summaries. This study employs ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE-1 measures the overlap at the unigram level, while ROUGE-2 calculates the overlap of two consecutive words (bigrams) to capture the the cohesion of semantic units and the structural continuity between sentence segments. ROUGE-L evaluates the similarity in sentence structure between the generated summary and the reference summary based on the Longest Common Subsequence (LCS). These metrics can effectively reflect the completeness and fluency of the information covered in the summary. It is one of the most common evaluation methods in summarisation tasks.

B 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! ;)

Summary from Raw Model *John has been working for his job.*

Summary from Anaphora-Resolved Model *Igor has two weeks left to finish work.*

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.*