Adapting Abstract Meaning Representation Parsing to the Clinical Narrative – the SPRING THYME parser

Jon Z. Cai[1], **Kristin Wright-Bettner**[1]
Martha Palmer[1], **Guergana K. Savova**[2], **James H. Martin**[1]
[1]University of Colorado Boulder
[2]Boston Children's Hospital and Harvard Medical School

Abstract

This paper is dedicated to the design and evaluation of the first AMR parser tailored for clinical notes. Our objective was to facilitate the precise transformation of the clinical notes into structured AMR expressions, thereby enhancing the interpretability and usability of clinical text data at scale. Leveraging the colon cancer dataset from the Temporal Histories of Your Medical Events (THYME) corpus, we adapted a state-of-the-art AMR parser utilizing continuous training. Our approach incorporates data augmentation techniques to enhance the accuracy of AMR structure predictions. Notably, through this learning strategy, our parser achieved an impressive F1 score of 88% on the THYME corpus's colon cancer dataset. Moreover, our research delved into the efficacy of data required for domain adaptation within the realm of clinical notes, presenting domain adaptation data requirements for AMR parsing. This exploration not only underscores the parser's robust performance but also highlights its potential in facilitating a deeper understanding of clinical narratives through structured semantic representations.

1 Introduction

Abstract Meaning Representation (Banarescu et al., 2013)(AMR)is a highly adaptable and expressive framework designed to capture the semantics of natural language expressions. Automatic AMR parsing is a natural language processing (NLP) method that translates natural language inputs into formal AMR expressions – representations which have proven to be useful across a wide range of downstream applications (Kapanipathi et al., 2021; Liu et al., 2015; Liao et al., 2018; Li and Flanigan, 2022; Bonial et al., 2020; Bai et al., 2021) including those in the biomedical domain (Garg et al., 2016; Rao et al., 2017).

Formally, AMR expressions take the form of labeled, rooted, directed, and acyclic graphs, $g = (V, E)$, where V represents the set of AMR nodes, which can be of type predicate, abstract concept and attributes; E represents the possible semantic relations between nodes such as prototypical agent and patient denoted by `arg0` and `arg1`. The AMR graph structure underpinned by Neo-Davidsonian semantics can then effectively encapsulate the abstract concepts, relationships, and entities present in individual sentences or utterances.

From a practical standpoint, AMR expressions encompass the semantic content typically addressed by individual representation schemes such as semantic role labeling (Palmer et al., 2005), named entities (Wang et al., 2022), and coreference chains (Joshi et al., 2020), thereby unifying these diverse aspects of meaning into a single comprehensive representation. Figure 1 illustrates an AMR expression selected from the clinical domain.

As Figure 1 demonstrates, concepts including events, entities and properties are captured as nodes in the graph, while the relations among the concepts are captured by labeled edges connecting the nodes. Events are represented using PropBank frames (Palmer et al., 2005), and the semantic relations of both entities and events to these predicates are specified either by a frame's numbered argument or one of the relations from AMR's role inventory. For example, the see-09 predicate represents the event of "visit/consultation by a medical professional." In this case, the agent of the seeing event is "Dr. Chandler Bing", represented by `see-09`'s `ARG0` relation, and the semantic role of patient for the event is "she" indicated by the `ARG1` semantic relation. AMR graphs also specify the temporal information in a formal way. In the above example, the time of the seeing event is specified by two temporal modifier subgraphs. It is a conjunction of "after now" and "within this week" which makes "later this week" a concrete time range.

AMR parsers based on pretrained large lan-

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 271–282
June 21, 2024 ©2024 Association for Computational Linguistics

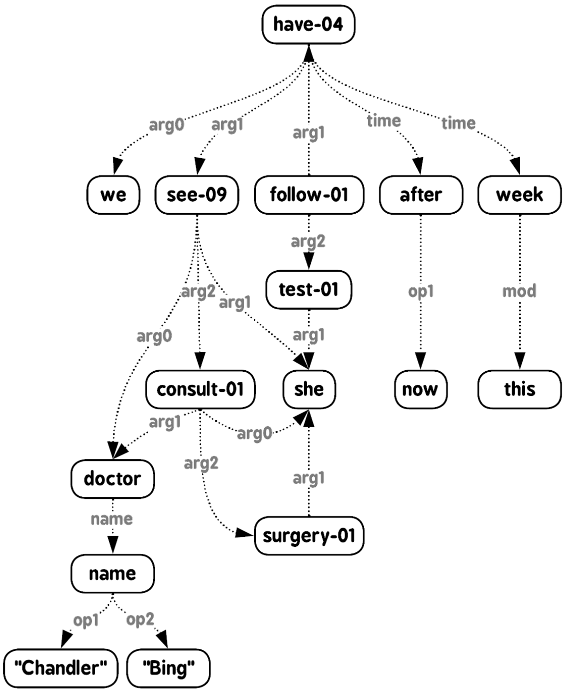

Figure 1: the AMR graph of sentence "We will have her see Dr. Chandler Bing in surgical consultation later this week following her testing."

guage models and sequence-to-sequence (encoder-decoder) architectures have demonstrated impressive accuracy when trained and evaluated on standard datasets. The use of AMR parsers has contributed to improved performance across a range of NLP tasks including question answering (Fu et al., 2021), information retrieval (Liao et al., 2018), knowledge-graph construction (Ribeiro et al., 2022), and text generation (Bai et al., 2022).

These successes have sparked growing interest in employing AMR in domains that diverge from the existing training data, such as human-robot interaction tasks, educational applications involving classroom discourse analysis, and diverse biomedical use cases. Unfortunately, as language form and meaning deviate from the general language captured in generic training data, parsing performance shows a rapid decline. This decline stems from disparities in vocabulary, syntax, and overall discourse structure. Addressing these challenges necessitates dedicated human expert annotation efforts to create domain-specific AMR resources. However, such endeavors can be costly and time-consuming. Hence, the preference lies in maximizing the utilization of existing data and parsers and adapting them to new domains, rather than building entirely new systems from scratch.

The contributions of this paper include:

- We adapted the high-performance SPRING parser (Bevilacqua et al., 2021) to the clinical domain, specifically leveraging the Temporal Histories of Your Medical Events (THYME) corpus (Wright-Bettner et al., 2020), and achieved state-of-the-art performance in AMR parsing within this context..
- We demonstrated that by tailoring an existing general domain English neural AMR parser with a relatively modest amount of gold-standard in-domain data, we could attain significantly high accuracy.
- We showcased data augmentation techniques that effectively enhance the parser's robustness across different domains.

2 Data

Supervised training data for AMR parsers consists of pairs of linguistic expressions along with their associated human annotated gold-standard AMR expressions. The current standard dataset for AMR development is AMR 3.0 (Knight et al., 2020) available from the Linguistic Data Consortium as LDC2020T02. This general domain dataset is the basis for our baseline efforts prior to domain adaptation. AMR 3.0 consists of over 59k English expressions from a variety of broadcast conversations, newswire, weblogs, web discussion forums, fiction and web text. To facilitate evaluation and model comparison, AMR 3.0 is divided into standard training, development and test splits consisting of 55,635, 1,722, and 1,898 expressions respectively.

To adapt AMR to the clinical narrative, we developed 8,327 in-domain AMRs (separate paper with detailed description under review) on a subset of the THYME colon cancer corpus (Styler et al., 2014; Wright-Bettner et al., 2020). The colon cancer part of the THYME corpus consists of 594 de-identified physicians' notes for 198 patients with colon cancer. Each patient is represented by one pathology note and two clinical notes. The corpus has undergone several prior annotation efforts, including temporal and coreference annotation (Styler et al., 2014; Wright-Bettner et al., 2019, 2020) and entity tagging as defined by the Unified Medical Language System (UMLS (Bodenreider, 2004)). As part of our AMR annotation process, we adopted seven clinical-domain named entity

(NE) types (anatomical-site, clinical-attribute, devices, disease-disorder, medications-drugs, sign-symptom) from the UMLS project and relied heavily on the UMLS in classifying many AMR concepts.

Like other genre-specific AMR tasks (Bonial et al., 2019; Bonn et al., 2020), we found it necessary to modify the standard AMR annotation approach to support meaningful annotation of domain-unique linguistic phenomena. Two phenomena are pervasive in the clinical narrative. First, physician notes frequently drop eventive mentions when they are inferable by human readers. For example, "Declines tetanus" does not mean the patient declined having tetanus; they declined a tetanus immunization. We expanded AMR's guidelines to permit explicit rendering of certain implicit concepts like the immunization:

```
(d / decline-02
      :ARG1 (s / shot-13 :implicit +
            :ARG3 (d2 / disease-disorder :
    name (n / name :op1 "tetanus"))))
```

Second, like other specialized domains, clinical texts are rife with semantically dense noun phrases (NPs) (Grön et al., 2018). In AMR, NPs must be treated in one of two ways: Either all components are extracted and related (white marble = marble that is white), or they are analyzed as single units of meaning, i.e., NEs (White House). However, semantic compositionality exists on a spectrum (Nakov, 2013), and many specialized NPs in particular strain the adequacy of a binary approach. This can be seen even in simple clinical NPs: One annotator might decide "blood pressure" is a single, cohesive unit of meaning and annotate it as an NE, while another might decide "pressure" is an extractable property of "blood". To address this, we implemented a two-pass strategy: In the first pass, for NPs that fell under one of the clinical NE types mentioned above, an experienced annotator made these compositionality judgments and added each unique phrase to a searchable, phrasal NE Dictionary along with an AMR fragment that "defined" the compositionality for each phrase. Annotators then referenced the Dictionary when building the AMR graphs in the second pass. This approach supported consistency and speed of annotation.

Finally, the THYME corpus contains frequent repetition of many other multiword expressions and phrases. For extremely formulaic phrases, such as those found in Vital Signs sections (Height = 167.60 cm, e.g.), we implemented a template-filling script that deterministically produced the AMRs, again saving significant manual annotation time. Of the 8,327 AMRs, 1,640 were produced by this script; the rest were created manually. The final 8,327 THYME-AMR data are split into training, development and test sets randomly with 4,955, 1,641 and 1,731 sentence-AMR pairs, respectively. All of the model training is conducted on the training set of the AMR 3.0 and THYME AMR corpora. We show the Inter Annotator Agreement between three annotators on 107 THYME-AMRs in Table 1

Comparison	P	R	F1
gold vs annotator 1	0.93	0.93	0.93
gold vs annotator 2	0.93	0.93	0.93
annotator 1 vs annotator 2	0.91	0.90	0.90

Table 1: Smatch scores on 107 manuall THYME AMRs, representing three clinical notes

3 Methods

We treat the AMR parsing task as a supervised machine learning problem and train a parameterized model to map natural language expressions to their corresponding AMR graphs. Various model architectures and training methods and paradigms have been employed over the years (Flanigan et al., 2014; Foland and Martin, 2017; Lyu and Titov, 2018; Cai and Lam, 2019; Zhang et al., 2019; Wang et al., 2015; Ballesteros and Al-Onaizan, 2017; Fernandez Astudillo et al., 2020; Hoang et al., 2021), resulting in a continuous improvement in the state of the art on the general domain AMR dataset(i.e. AMR 2.0 and 3.0 corpus (LDC2020T2)). However, these improvements are highly dependent on the availability of significant amounts of annotated training data hampering the development of parsers for specific genres and languages other than English. Our approach here is to leverage an existing high-performance parser and adapt it to the clinical domain using the modest amount of domain-specific training data described in the last section.

Meanwhile, the great advances of the pre-trained foundational models has introduced a new modeling paradigm in the field of NLP as well as to structure-prediction problems such as AMR parsing. In particular, the sequence-to-sequence modeling, originally developed for machine translation, has proven a highly effective approach for AMR parsing (Bevilacqua et al., 2021; Konstas et al., 2017; Xu et al., 2020). In this approach, two neu-

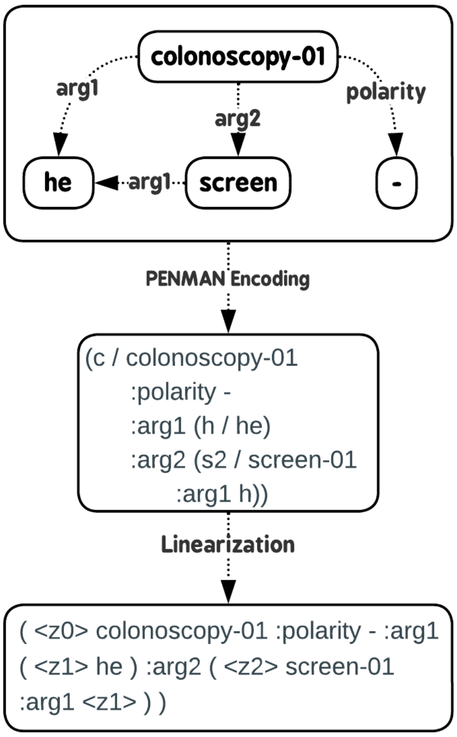

Figure 2: AMR graph to PENMAN linearization pipeline. The transformation map between the AMR graphical representation and its linearized representation is one-to-one-and-onto.

ral network components are involved: an encoder, which takes the natural language sentence as input and maps it to a continuous manifold as a sequence of high-dimensional vectors, and a decoder, which takes the embedded sentence representation vectors and maps them to the output embedding space, corresponding to the target sequence tokens.

Here we make use of the SPRING parser (Bevilacqua et al., 2021), one of the state-of-the-art AMR parsers on AMR 3.0 evaluation. The underlying pre-trained language model is BART-large (Lewis et al., 2020), a transformer-based language model that has been trained using a set of denoising pre-training objectives, such as a masked language modeling objective and a document reconstruction objective, on general domain unlabeled English text. The neural network architecture relies on the self-attention and cross-attention mechanism to learn patterns from natural language texts. This pre-trained model is then fine-tuned on the AMR 3.0 training data to map English inputs to linearized AMR graphs, which consist of a sequence of AMR tokens. We show the linearization correspondence of an AMR graph to its sequence of AMR tokens in Figure 2.

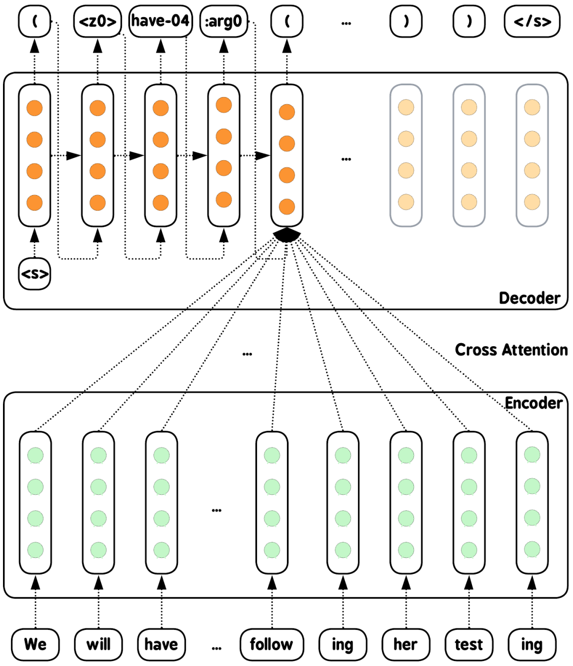

Figure 3: The SPRING parser modeling diagram. A transformer-based self-attention mechanism is used to produce embeddings for the input expression. The decoder then uses cross attention to drive autoregressive generation of a sequence of AMR output tokens.

A critical aspect of using sequence-to-sequence models for structured prediction tasks, like parsing, is transforming the task itself. In AMR parsing, the AMR graph is converted into a sequence of tokens through a linearization algorithm. Note that the vocabulary of the decoder differs from that of the encoder model, as the target sequence consists of AMR-specific tokens such as the relations `arg0` and `arg1`, and predicates like `test-01`. During fine-tuning, we utilize the vocabulary derived from the AMR 3.0 corpus, which ensures consistency and accuracy in the parsing process. The parsing problem is then to convert an input text sequence into a valid sequence of AMR tokens that can be deterministically transformed into a directed AMR graph. The overall SPRING approach is depicted in Figure 3. Given a high-performing SPRING model, we adapt it to the THYME domain by fine-tuning on the THYME-AMR training set (4,955 expressions). Here, fine-tuning involves continuous gradient-based updates to the original model parameters with a small learning rate (5×10^{-6}) with batch size to be 20, we keep the maximum sequence length to be 1024..

3.1 Evaluation

The standard metric to evaluate AMR parsing performance is SMATCH, which decomposes an AMR graph into triples that capture the edge list representation of a graph structure. For instance, the AMR for the sentence “He had never undergone a screening colonoscopy.” can be decomposed into its edge list representation as AMR1 and edge list 1 as follows:

```
AMR1:
(c / colonoscopy-01 :polarity -
      :arg1 (h / he)
      :arg2 (s2 / screen-01
            :arg1 h))

AMR2:
(c1 / colonoscopy-01 :polarity -
      :arg1 (s / she)
      :arg2 (s2 / screen-01
            :arg1 s))
```

```
Decomposed edge list1:
      instance(c, colonoscopy-01)
      instance(h, he)
      instance(s2, screen-01)
      polarity(c, -)
      arg1(c, h)
      arg2(c, s2)
      arg1(s2, h)

Decomposed edge list2:
      instance(c1, colonoscopy-01)
      instance(s, she)
      instance(s2, screen-01)
      polarity(c1, -)
      arg1(c1, s)
      arg2(c1, s2)
      arg1(s2, s)
```

We conjured another slightly altered AMR2 with the `he` node replaced with a `she` node, indicating a potential mistake in the parser generated AMR. In the above decomposition of AMR graphs, `instance()` represents the nodes in the graph while the rest are the edges. Given the edge lists for a hypothetical parse and its corresponding gold-standard parse, the SMATCH metric produces precision (p), recall (r), and F1-measure scores as follows:

$$p = \frac{N_{correct}}{N_{predicted}}, r = \frac{N_{correct}}{N_{reference}}, F_1 = \frac{2pr}{p+r}$$

A complication in computing these scores is that we need to know which of the proposed AMR nodes in the parse are supposed to correspond to which ones in the correct set. In other words, the graphs need to be matched before they can be scored. This issue originates from the encoding of AMR nodes with variables, through which different instantiations of a concept can be encoded. The standard SMATCH scorer (Cai and Knight, 2013) employs a greedy heuristic method to provide the required alignment to avoid computing a computationally expensive optimal alignment.

Finally, AMR representations are an amalgamation of semantic representations including predicate-argument relations, named entities, and coreference components. The SMATCH score represents an average over these component categories, obscuring the model performance over the various categories of information in AMR expressions, thus making it difficult to assess the usability of the results in downstream applications. To address this, a more fine-grained analysis tool[1] provides precision, recall and F1 measures across the various component AMR tasks. We will discuss the fine-grained categories in section 4.3.

4 Experiments

We present the domain adaption training experiments in this section to show the characteristics of the text from THYME corpus when it comes to AMR parser developement.

4.1 Domain Adaptation

Table 2 provides the results of our primary domain adaptation experiments. The first column presents the evaluation results of the off-the-shelf SPRING AMR parser trained solely with the AMR 3.0 training data. The 83.0 SMATCH score for the SPRING parser reaches near state-of-the-art performance on the AMR 3.0 test set, whereas, the performance on the THYME-AMR test set is significantly lower at 51.7 SMATCH. The second column shows the results of the same parser fine-tuned using the THYME-AMR training data. Here, we see that the fine-tuned parser achieves excellent results on the THYME-AMR corpus test set with a 35.3 point absolute improvement over the original model.

Train Test	AMR 3.0	THYME-AMR	AMR 3.0 + THYME-AMR
AMR 3.0	83.0	77.0	80.0
THYME-AMR	51.7	87.0	88.0

Table 2: SPRING THYME-AMR parser performance with different training sources. All scores are Smatch F1

[1] https://github.com/mdtux89/amr-evaluation

4.2 Avoiding Forgetting

Catastrophic forgetting is a frequently observed problem when fine-tuning large pre-trained models on domain specific data (Li and Hoiem, 2018; Riemer et al., 2019; Scialom et al., 2022). While fitting the model's parameters to the new domain, there is often a significant loss in terms of the model's performance on its original domain. To assess the robustness and potential forgetting of general domain AMR knowledge, we evaluated the THYME-AMR fine-tuned parser on the AMR 3.0. The results showed a decrease in performance from 83.8 to 77.0, indicating significant forgetting of the general domain AMR.

Based on this observation, we deployed a joint training approach to mitigate this forgetting phenomenon. In this experiment, we fine-tuned the parser on a mixture sampled from both the AMR 3.0 and THYME-AMR data. Considering the differing sizes of the two corpora, we sampled them in a 12-to-1 ratio between THYME-AMR and AMR 3.0 sources. As can be seen from Table 2, this modest infusion of general domain data allowed the parser to attain high performance on the THYME-AMR test set while also largely maintaining its performance on the AMR 3.0 test set. This observation underscores the effectiveness of domain-specific annotation in improving semantic parsing in a joint fashion. This means that the understanding of semantics improves collectively rather than independently, thanks to domain-specific data. As more representative data are collected, we expect further improvements in the parser's performance, making it even more adept at comprehending the semantics in the given domain.

4.3 Fine-Grained Performance

Table 3 presents detailed results of our best-performing parser across the semantic components that comprise AMR graphs. AMR representations are an amalgamation of semantic representations including predicate-argument relations, named entities, and coreference components. The SMATCH score represents an average over these sub-categories. To leverage the in-depth analytical power of these linguistic sub categories, a more fine-grained analysis tool[2] provides precision, recall and F1 measures across the various component AMR tasks. We list the fine-grained performance metric category definitions briefly as follows:

- *Unlabeled* category assesses the parsing performance on the AMR graph, disregarding the edge labels.

- *No WSD* category evaluates the parsing performance while ignoring the Propbank word sense labels (e.g., `see-09` becomes just `see`).

- *Concepts* category considers only the abstract concept node matches.

- *Named Entity* category focuses on the matches of named entity subgraphs.

- *Negation* category concerns the matches of the negation attribute nodes(e.g. the `:polarity` edges).

- *Reentrancy* category examines only the concept re-entrancy subgraphs(usually a back reference node).

- *Semantic Role Label (SRL)* category pertains to the performance of each predicate argument structure generation.

We observe that the mixed data augmentation technique significantly improves performance across the board, impacting almost every subcategory of evaluation. Notably, the off-the-shelf parser faced significant challenges in understanding the semantics in the new domain. The performance drop due to domain shifting was not uniform across different sub-categories. The most significant drop in performance was seen in *Named Entity* Recognition, which is expected due to the abundance of medical-related terminology. On the other hand, the data-augmented parser excelled in *Concept* predication and *Named Entity* recognition aspects of AMR parsing, while the performance in the *Negation* and *Reentrancy* category was relatively less impressive compared to the other categories.

4.4 Data Requirements for Successful Adaptation

Manual annotation of AMR data is time consuming and expensive. At the current time, the standard AMR 3.0 still consists of only 60k sentences, nearly 10 years after the initial data release. The results shown in Table 2 raise the question of how

[2] `https://github.com/mdtux89/amr-evaluation`

Sub-category	Training Set	Precision	Recall	F1
SMATCH	THYME-AMR + AMR 3.0	0.89	0.88	0.88
	THYME-AMR	0.88	0.87	0.87
	AMR 3.0	0.53	0.45	0.49
Unlabeled	THYME-AMR + AMR 3.0	0.90	0.90	0.90
	THYME-AMR	0.90	0.88	0.89
	AMR 3.0	0.60	0.51	0.55
No WSD	THYME-AMR + AMR 3.0	0.89	0.88	0.88
	THYME-AMR	0.88	0.87	0.87
	AMR 3.0	0.55	0.46	0.50
Concepts	THYME-AMR + AMR 3.0	0.93	0.92	0.93
	THYME-AMR	0.93	0.91	0.92
	AMR 3.0	0.52	0.46	0.49
Named Ent.	THYME-AMR + AMR 3.0	0.94	0.93	0.93
	THYME-AMR	0.93	0.92	0.92
	AMR 3.0	0.18	0.05	0.08
Negation	THYME-AMR + AMR 3.0	0.86	0.85	0.85
	THYME-AMR	0.84	0.86	0.85
	AMR 3.0	0.45	0.42	0.44
Reentrancies	THYME-AMR + AMR 3.0	0.78	0.79	0.78
	THYME-AMR	0.78	0.76	0.77
	AMR 3.0	0.48	0.37	0.41
SRL	THYME-AMR + AMR 3.0	0.88	0.87	0.87
	THYME-AMR	0.87	0.85	0.86
	AMR 3.0	0.55	0.47	0.51

Table 3: SPRING parser performance analytical breakdowns comparison among three models trained on different combination of the fine-tuning data source. The evaluation is on the THYME-AMR test set.

much data is actually required to attain high levels of parser accuracy through adaptation. To address this question, we conducted a series of experiments training models with progressively larger snapshots of the available training data. Specifically, we gradually augmented the training set size for each model by random sampling without replacement from the training data (resulting in training sets of size 500, 1,000, 2,000, 3,000, 4,000 and 4,955). The results in Figure 4 illustrate the parser's performance across these training sets.

As can be seen, performance rapidly rises from the non-adapted baseline to 80 SMATCH with 1,000 training examples; the model trained on only 2,000 samples achieves 90% of the performance of our best parser trained on all available training data. This rapid improvement with domain specific data is a positive indication of the effectiveness of continued training from a generic model and its ability to rapidly generalize from the domain-specific data.

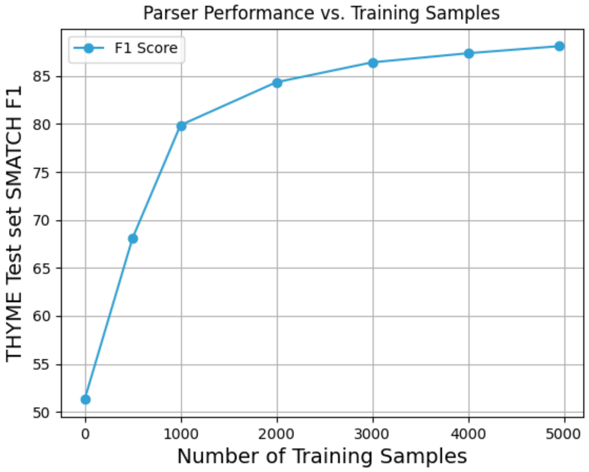

Figure 4: The performance curve with different sample sizes of the THYME-AMR training set. The x axis is the sample size of the training data; the y axis represents the SMATCH F1 performance score(with the unit of percentage) of the parsers evaluated on the same withheld test set (THYME-AMR test set)

5 Discussion

Our results have highlighted the advantages of employing data augmentation techniques for domain adaptation fine-tuning. This opens up the possibility for additional follow-up studies, including the incorporation of data from domain-specific Propbank roleset development. For instance, in the case of THYME, leveraging example sentences for newly added named-entity types like "anatomical-site" could prove beneficial. Initializing the word embedding vectors with such domain-specific concepts would enable a better fit with the pre-trained foundational models. Future investigations involving more sophisticated foundational models and data augmentation approaches hold great promise for enhancing AMR parsing in the medical domain and other specialized domains. By harnessing the capabilities of cutting-edge language models and innovative data augmentation strategies, we can expect significant advancements in semantic parsing tasks and domain adaptation techniques.

With these advances, AMR parses have wide applicability to core information extraction tasks from the clinical narrative such as entity recognition, negation detection, uncertainty detection, coreference, temporality and relation extraction.

6 Conclusion

In our investigation, we have presented substantial evidence highlighting the critical role of domain-specific AMR annotations in the context of domain adaptation. Our findings illuminate how variances in the distribution between original and target domains can precipitate a marked decline in the performance of AMR parsing. This phenomenon underscores the challenge of catastrophic forgetting, a significant hurdle in the training of neural network models where new learning can disrupt previously acquired knowledge.

To counteract this issue, we demonstrated the critical role of data augmentation techniques. Specifically, by integrating domain-specific examples into the training dataset, we significantly bolstered the model's capability to acclimate to the nuances of the new domain while preserving its proficiency in the original domain. This strategic approach of coupling domain-specific annotation with thoughtful data augmentation has emerged as a formidable solution, ensuring both the robustness and accuracy of AMR parsing across different domain adaptation scenarios.

Our study reaffirms the indispensability of domain-specific annotation in achieving effective domain adaptation and also supports data augmentation as an essential tool in maintaining a delicate balance between learning new domain characteristics and retaining essential knowledge from the original domain. This balanced approach provides a promising avenue for future research and development in the field of AMR parsing, potentially paving the way for more nuanced and adaptable AI systems capable of navigating other domains with limited data yet maintain robustness.

7 Limitations and Future Work

Our study faced constraints primarily due to computational limitations, which necessitated a focus on a specific subset of model and data augmentation strategies. A reasonable extension of this research could involve the exploration of more advanced foundational models, including GPT-3.5, GPT-4, and their publicly accessible counterparts such as LLAMA. These platforms present opportunities for experimenting with zero- or few-shot learning techniques. Importantly, our use of clinical data mandates adherence to stringent privacy standards; thus, it is imperative that any models employed can be locally installed and operated within a secure, firewall-protected environment. This requirement currently excludes the use of proprietary models like those within the GPT family, which are tailored for commercial applications and do not meet the privacy criteria essential for our research objectives.

8 Acknowledgements

This research and the paper it informs have been funded by the United States National Institutes of Health (grants R01LM010090, R01LM013486). The content is solely the responsibility of the authors and does not necessarily represent the official views of the United States National Institutes of Health. The results presented in this paper were obtained using the Chameleon testbed(Keahey et al., 2020) supported by the National Science Foundation. We wish to extend our heartfelt gratitude to Ahmed Elsayed and Gaby Dinh for their exceptional annotation efforts, without which this research would not have been possible. Additionally, we are deeply appreciative of the thorough discussions and invaluable suggestions from Skatje Myers, Steven Bethard, David Harris, Timothy Miller,

Danielle Bitterman, Piet de Groen, and Dmitriy Dligach.

Ethics Statement

In our exploration of clinical notes analysis and the design of automation systems, we navigate through a terrain rich with sensitive personal data and entwined with ethical complexities. Our work is fundamentally rooted in a profound respect for the dignity, rights, and welfare of the individuals whose lives and experiences are documented in these notes. Guided by a set of core ethical principles, our research endeavors to uphold the highest standards of integrity and respect.

Foremost, we prioritize the privacy and confidentiality of patient data. In this paper, all examples have been rigorously de-identified to ensure no personal information can lead back to individuals. Moreover, recognizing the critical importance of obtaining informed consent, we actively collaborate with institutional review boards (IRB) to ethically justify and secure consent approvals for utilizing all data involved in our research.

We are acutely aware of the potential biases in our analysis and interpretation of clinical narratives. This awareness extends to biases that might emerge from the data collection process, the selection of narratives for analysis, and our own preconceptions. We are committed to making concerted efforts to ensure that our analysis encompasses diverse perspectives, thereby avoiding the perpetuation of stereotypes or inequalities.

We urge downstream users of our parser to conscientiously consider the potential impact of their findings on the individuals depicted in the clinical narratives, as well as on wider patient populations. This involves thoughtful reflection on how the research could affect public perceptions, clinical practice, and policy making. A crucial aspect of our approach is to balance the dissemination of research findings with the imperative to prevent harm or distress.

Lastly, our pursuit of transparency in our methodology and findings is relentless. We advocate for the use of Abstract Meaning Representation (AMR) as a superior tool compared to opaque, “black-box” models. AMR offers a fully transparent and verifiable representation of the semantics in clinical narratives, which aligns with our commitment to fostering trust and accountability.

Our approach is a testament to our dedication to ethical research practices, emphasizing the protection of privacy, the mitigation of bias, the thoughtful consideration of impacts, and the advancement of transparency and accountability. These principles are the bedrock of our efforts to contribute meaningful and ethically sound advancements in the field of clinical notes analysis and automation system design.

References

Xuefeng Bai, Yulong Chen, Linfeng Song, and Yue Zhang. 2021. Semantic representation for dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4430–4445, Online. Association for Computational Linguistics.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

Miguel Ballesteros and Yaser Al-Onaizan. 2017. AMR parsing using stack-LSTMs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1269–1275, Copenhagen, Denmark. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Proceedings of AAAI*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-AMR: Abstract Meaning Representation for dialogue. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 684–695, Marseille, France. European Language Resources Association.

Claire N. Bonial, Lucia Donatelli, Jessica Ervin, and Clare R. Voss. 2019. Abstract Meaning Representation for human-robot dialogue. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 236–246.

Julia Bonn, Martha Palmer, Zheng Cai, and Kristin Wright-Bettner. 2020. Spatial AMR: Expanded spatial annotation in the context of a grounded Minecraft corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4883–4892, Marseille, France. European Language Resources Association.

Deng Cai and Wai Lam. 2019. Core semantic first: A top-down approach for AMR parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3799–3809, Hong Kong, China. Association for Computational Linguistics.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Ramón Fernandez Astudillo, Miguel Ballesteros, Tahira Naseem, Austin Blodgett, and Radu Florian. 2020. Transition-based parsing with stack-transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1001–1007, Online. Association for Computational Linguistics.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the Abstract Meaning Representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Baltimore, Maryland. Association for Computational Linguistics.

William Foland and James H. Martin. 2017. Abstract Meaning Representation parsing using LSTM recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–472, Vancouver, Canada. Association for Computational Linguistics.

Ruiliu Fu, Han Wang, Xuejun Zhang, Jun Zhou, and Yonghong Yan. 2021. Decomposing complex questions makes multi-hop QA easier and more interpretable. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 169–180, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sahil Garg, Aram Galstyan, Ulf Hermjakob, and Daniel Marcu. 2016. Extracting biomolecular interactions using semantic parsing of biomedical text. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2718–2726. AAAI Press.

Leonie Grön, Ann Bertels, and Kris Heylen. 2018. The interplay of form and meaning in complex medical terms: Evidence from a clinical corpus. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 18–29, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Thanh Lam Hoang, Gabriele Picco, Yufang Hou, Young-Suk Lee, Lam M. Nguyen, Dzung T. Phan, Vanessa López, and Ramon Fernandez Astudillo. 2021. Ensembling graph predictions for AMR parsing. In *Advances in Neural Information Processing Systems*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramón Fernandez Astudillo, Maria Chang, Cristina Cornelio, Saswati Dana, Achille Fokoue, Dinesh Garg, Alfio Gliozzo, Sairam Gurajada, Hima Karanam, Naweed Khan, Dinesh Khandelwal, Young-Suk Lee, Yunyao Li, Francois Luus, Ndivhuwo Makondo, Nandana Mihindukulasooriya, Tahira Naseem, Sumit Neelam, Lucian Popa, Revanth Gangi Reddy, Ryan Riegel, Gaetano Rossiello, Udit Sharma, G P Shrivatsa Bhargav, and Mo Yu. 2021. Leveraging Abstract Meaning Representation for knowledge base question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3884–3894, Online. Association for Computational Linguistics.

Kate Keahey, Jason Anderson, Zhuo Zhen, Pierre Riteau, Paul Ruth, Dan Stanzione, Mert Cevik, Jacob Colleran, Haryadi S. Gunawi, Cody Hammock, Joe Mambretti, Alexander Barnes, François Halbach, Alex Rocha, and Joe Stubbs. 2020. Lessons learned from the chameleon testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association.

Kevin Knight, Bianca Badarau, Laura Baranescu, Claire Bonial, Madalina Bardocz, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, Tim O'Gorman, and Nathan Schneider. 2020. Abstract meaning representation (amr) annotation release 3.0. Web Download. LDC Catalog No.: LDC2020T02, DOI: `https://doi.org/10.35111/44cy-bp51`.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of* *the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Changmao Li and Jeffrey Flanigan. 2022. Improving neural machine translation with the Abstract Meaning Representation by combining graph and sequence transformers. In *Proceedings of the 2nd Workshop on Deep Learning on Graphs for Natural Language Processing (DLG4NLP 2022)*, pages 12–21, Seattle, Washington. Association for Computational Linguistics.

Zhizhong Li and Derek Hoiem. 2018. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract Meaning Representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1178–1190, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A. Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Denver, Colorado. Association for Computational Linguistics.

Chunchuan Lyu and Ivan Titov. 2018. AMR parsing as graph prediction with latent alignment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 397–407, Melbourne, Australia. Association for Computational Linguistics.

Preslav Nakov. 2013. On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, 19:291 – 330.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Sudha Rao, Daniel Marcu, Kevin Knight, and Hal Daumé III. 2017. Biomedical event extraction using Abstract Meaning Representation. In *BioNLP 2017*, pages 126–135, Vancouver, Canada,. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, , and Gerald Tesauro. 2019. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

IV Styler, William F., Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015. A transition-based algorithm for AMR parsing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375, Denver, Colorado. Association for Computational Linguistics.

Yu Wang, Hanghang Tong, Ziye Zhu, and Yun Li. 2022. Nested named entity recognition: A survey. *ACM Trans. Knowl. Discov. Data*, 16(6).

Kristin Wright-Bettner, Chen Lin, Timothy Miller, Steven Bethard, Dmitriy Dligach, Martha Palmer, James H. Martin, and Guergana Savova. 2020. Defining and learning refined temporal relations in the clinical narrative. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 104–114, Online. Association for Computational Linguistics.

Kristin Wright-Bettner, Martha Palmer, Guergana Savova, Piet de Groen, and Timothy Miller. 2019. Cross-document coreference: An approach to capturing coreference without context. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 1–10, Hong Kong. Association for Computational Linguistics.

Dongqin Xu, Junhui Li, Muhua Zhu, Min Zhang, and Guodong Zhou. 2020. Improving AMR parsing with sequence-to-sequence pre-training. In *Proceedings* *of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2501–2511, Online. Association for Computational Linguistics.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019. AMR parsing as sequence-to-graph transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 80–94, Florence, Italy. Association for Computational Linguistics.