

# Extracting Appointment Spans from Medical Conversations

**Nimshi Venkat Meripo**  
Abridge AI Inc.  
venkatm@abridge.com

**Sandeep Konam**  
Abridge AI Inc.  
san@abridge.com

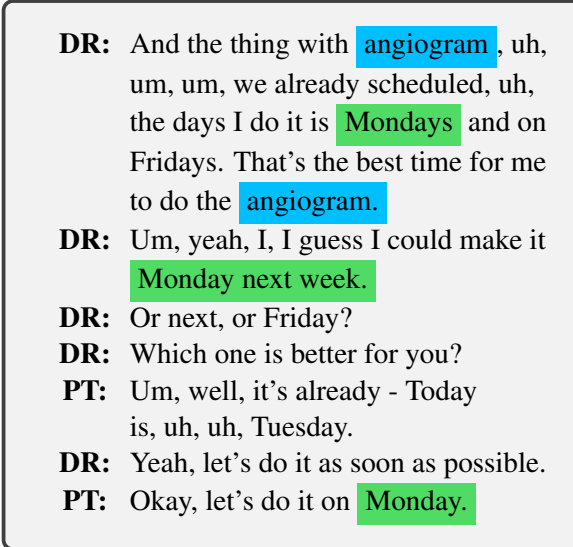
## Abstract

Extracting structured information from medical conversations can reduce the documentation burden for doctors and help patients follow through with their care plan. In this paper, we introduce a novel task of extracting appointment spans from medical conversations. We frame this task as a sequence tagging problem and focus on extracting spans for appointment reason and time. However, annotating medical conversations is expensive, time-consuming, and requires considerable domain expertise. Hence, we propose to leverage weak supervision approaches, namely incomplete supervision, inaccurate supervision, and a hybrid supervision approach and evaluate both generic and domain-specific, ELMo, and BERT embeddings using sequence tagging models. The best performing model is the domain-specific BERT variant using weak hybrid supervision and obtains an F1 score of 79.32.

## 1 Introduction

Increased Electronic Health Records (EHR) documentation burden is one of the leading causes for physician burnout (Downing et al., 2018; Collier, 2017). Although EHRs facilitate effective workflow and access to data, several studies have shown that physicians spend more than half of their workday on EHRs (Arndt et al., 2017). This leads to decreased face time with patients and reduced work satisfaction for physicians (Drossman and Ruddy, 2019; Sinsky et al., 2016). For these reasons, there has been growing interest in using machine learning techniques to extract relevant information for a medical record from medical conversations (Lin et al., 2018; Schloss and Konam, 2020).

On the other hand, research shows that approximately 23% of patients do not show up for their doctor appointments (Dantas et al., 2018). Missed appointments have a large impact on hospitals' ability to provide efficient and effective services (Chandio et al., 2018). Studies in Callen et al. (2012)



**DR:** And the thing with **angiogram**, uh, um, um, we already scheduled, uh, the days I do it is **Mondays** and on Fridays. That's the best time for me to do the **angiogram**.  
**DR:** Um, yeah, I, I guess I could make it **Monday next week**.  
**DR:** Or next, or Friday?  
**DR:** Which one is better for you?  
**PT:** Um, well, it's already - Today is, uh, uh, Tuesday.  
**DR:** Yeah, let's do it as soon as possible.  
**PT:** Okay, let's do it on **Monday**.

Figure 1: An utterance window from a medical conversation annotated with appointment **reason** and **time** spans.

also show that a significant number of patients miss their lab appointments. Missed lab appointments can put a patient's health at risk and allow diseases to progress unnoticed (Mookadam et al., 2016). One of the main reasons for no-shows is patient forgetfulness (Ullah et al., 2018). Mookadam et al. (2016) and Perron et al. (2013) show that proactive reminders through text messages, calls, and mobile applications are promising and significantly decrease the missed appointment rates.

In line with the aforementioned value, appointment span extraction from medical conversations can help physicians document the care plan regarding diagnostics (Dx), procedures (Px), follow-ups, and referrals. It can also directly impact a patient's ability to keep their appointments. In this work, we investigate extracting the appointment reason and time spans from medical conversations as shown in Figure 1. The reason span refers to a phrase that corresponds to Dx, Px, follow-ups and referrals.

The time span refers to a phrase that corresponds to the time of the appointment. To tackle this task, we collected a dataset for both reason and time spans and framed it as a sequence tagging problem. Our contributions include: (i) defining the appointment span extraction task, (ii) describing the annotation methodology for labeling the medical conversations, (iii) investigating weak supervision approaches on sequence tagging models using both generic and domain-specific ELMo and BERT embeddings, and (iv) performing error analysis to gather insights for improving the performance.

## 2 Related work

Extracting Dx, Px and time expressions has been the subject of past work. Tools such as Clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova et al., 2010) and MetaMa (Aronson, 2006) are widely used in the biomedical field to extract medical entities. Both of these tools use the Unified Medical Language System (UMLS) (Bodenreider, 2004) to extract and standardize medical concepts. However, UMLS is primarily designed for written clinical text, not for spoken medical conversations. Further, research on date-time entity extraction from text is task agnostic. Rule-based approaches like HeidelTime (Strötgen and Gertz, 2010), and SUTime (Chang and Manning, 2012) mainly handcraft rules to identify time expression in the text. Learning-based approaches typically extract features from text and apply statistical models such as Conditional Random Fields (CRFs). While these tools perform well for generic clinical and date-time entity extraction from texts, they don’t fare as well on task-specific entity extraction, where only a subset of the entities present in the text is relevant to solving the task.

Recently, there has been an increasing interest in medical conversations-centered applications (Chiu et al., 2017). Du et al. (2019a,b) proposed several methods for extracting entities such as symptoms and medications and their relations. Selvaraj and Konam (2019) and Patel et al. (2020) examined the task of medication regimen extraction. While recent research in medical conversations is primarily focused on extracting symptoms and medications, we propose a new task of extracting appointment spans. Our framing of this task as a sequence tagging problem is similar to Du et al. (2019a,b); however, they use a fully supervised approach and mainly focus on relation extraction, whereas we

investigate weak supervision for appointment span extraction. Moreover, we evaluate both generic and domain-specific ELMo and BERT models in our task.

## 3 The Appointment Span Extraction Task

### 3.1 Corpus Description

Our corpus consists of human-written transcripts of 23k fully-consented and manually de-identified real doctor-patient conversations. Each transcript is annotated with utterance windows where the appointment is discussed. We have obtained a total of 43k utterance windows that discuss appointments. Of the 43k utterance windows, 3.2k utterances windows from 5k conversations are annotated with two types of spans: appointment reason and appointment time (Figure 1). We have also obtained annotations for other span types such as appointment duration and frequency, however due to infrequency of such spans, we have not included these spans in this study.

### 3.2 Annotation Methodology

| Span Type | Examples  |
|-----------|---|
| Reason    | follow-up, dermatologist, MRI, chemotherapy, chemo, physical, heart surgery                     |
| Time      | about a month, every two weeks, in the middle of August, July 2021, before the next appointment |

Table 1: Examples of annotated spans.

A team of 15 annotators annotated the dataset. The annotators were highly familiar with medical language and have significant experience in medical transcription and billing. We have distributed 3.2k utterance windows equally among 15 annotators. Each utterance window is doubly-annotated with appointment spans, and the authors resolved any conflicting annotations. We collect the spans of text describing the reason and time for only future appointments. We show examples of reason and time spans in Table 1. Overall, 6860 reason spans and 2012 time spans are annotated, and the average word lengths for reason and time spans are 1.6 and 2.3, respectively.

**Reason span** The reason span captures four types of appointments: follow-ups, referrals, diagnostics, and procedures. Phrases of body parts and substances are also captured if they are mentioned in relation to the appointment reason (e.g., ultrasound of my *kidney*, surgery for the *heart valve*). We also annotated the spans where the appointment reason is expressed in informal language (e.g., *see you back* for follow-ups, let’s do your *blood* for a blood test).

**Time span** The time span captures the time of an appointment. We also included prepositions (eg. *in* two days, *at* 3 o’clock) and time modifiers (eg., *after* a week, *every* year) in this span. In cases where multiple different time phrases are present for an appointment, annotators were instructed to annotate a time phrase that is confirmed by either patient or doctor, or annotate potentially valid time phrases if the discussion is ambiguous.

Due to the conversational nature, appointment reason and time are often discussed multiple times using the same phrase or a synonymous phrase (e.g., a *blood test* called *FibroTest*, *Monday* or *Monday next week*). To maintain consistency across different conversations, annotators were instructed to mark all occurrences of the span.

### 3.3 Methods

To account for the limited set of annotations, we employed weak supervision approaches. We specifically used inaccurate supervision, incomplete supervision (Zhou, 2018) and developed a hybrid approach that utilizes both inaccurate and incomplete supervision.

**Inaccurate Supervision** Inaccurate supervision is a scenario where the training labels provided are not always the ground-truth; in other words, the training labels suffer from errors. We take advantage of off-the-shelf tools such as UMLS and spaCy (Honnibal et al., 2020) to automatically annotate reason and time spans. For the reason span, we perform a dictionary lookup in UMLS vocabularies and extract any span with a semantic type belonging to Dx, Px, and body parts. For the time span, we use spaCy’s named entity recognition (NER) model to extract spans belonging to time and date. To reduce the inaccuracies, we included only the utterance windows with at least one reason phrase and one time phrase. Using this approach, we ob-

tained 20k utterance windows with both appointment reason and time spans.

**Incomplete Supervision** Incomplete Supervision refers to a scenario where only a small subset of data has annotated labels. For this scenario, we use 2.5k conversations from manual span annotated corpus conversations, which resulted in 1292 utterance windows.

**Hybrid Supervision** In this approach, we apply both inaccurate and incomplete supervision techniques sequentially. To avoid catastrophic forgetting (McCloskey and Cohen, 1989), the models are first trained with inaccurate supervision and then fine-tuned with incomplete supervision.

We use a 85:15 split of the remaining 1844 manual span annotated utterance windows for testing and validation purposes. To make the test dataset more difficult, we used a weighted sampling technique in which each appointment span is weighted by the inverse probability of it being sampled.

## 4 Models

In this section, we briefly describe our two models that use variants of contextualized embeddings namely, ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018).

### 4.1 ELMo-based CRF

Our model is a 2-layer BiLSTM network using GloVe word embeddings and a character-based CNN representation trained with CRF loss. Similar to the approach taken in Peters et al. (2018), our model is enhanced by concatenating a weighted average of ELMo embeddings with GloVe and character embeddings. We next describe the two variants of ELMo models we use.

**ELMo** The original ELMo model is pre-trained on generic language corpora using the 1-Billion Words dataset (Chelba et al., 2013).

**BioELMo** BioELMo (Jin et al., 2019) is a biomedical variant of ELMo trained on 10M recent abstracts (2.46B tokens) from PubMed.

### 4.2 BERT-based classifier

Similar to the approach taken in Devlin et al. (2018), we use a token level classifier instead of a CRF layer and fine-tune variants of the BERT model. We next describe the variants of BERT models we use.

**BERT** The original BERT model is trained on BooksCorpus (Zhu et al., 2015) and English Wiki.

**BioBERT** BioBERT (Lee et al., 2019) further pre-trains the BERT-base model on a large corpus of PubMed abstracts containing 4.5B words.

### 4.3 Experiment details

| Model         | Embedding Size | Learning rate |
|---------------|----------------|---------------|
| ELMo variants | 1024           | 1e-3          |
| BERT variants | 768            | 3e-5          |

Table 2: Experiment configurations for the models.

The experiment configuration for ELMo and BERT variants used in our experiments is shown in Table 2. Both ELMo and BERT variants use an uncased vocabulary. The span labels are represented using the IOB2 tagging scheme (Sang and Veenstra, 1999).

## 5 Evaluation

To evaluate our models, we measure micro-averaged Precision, Recall, and F1 of reason and time spans on the test dataset (Table 3). Both ELMo and BERT variants performed similarly with inaccurate supervision owing to the noisy nature of the inaccurate supervision. With the incomplete supervision approach, the performance improved considerably, ranging from 49% in ELMo to 60% in BioBERT. Both BioELMo and BioBERT gained more than the ELMo and BERT variants, respectively. However, with hybrid supervision, both the ELMo variants benefited most and achieved similar performance nullifying the advantage of the in-domain pre-training of BioELMo.

On the other hand, the BERT variants showed a minor improvement with hybrid supervision. The BERT variants consistently performed better than ELMo variants, and the domain-specific pre-training has only a minor impact on BERT when compared to ELMo. Overall, the proposed hybrid supervision approach has consistently improved performance across all model variants and the results show that augmenting the training data with inaccurate supervision can improve the performance.

In order to assess performance at each span type, we chose the best performing BioBERT-hybrid model. For both span types precision was lower than recall (Table 4) suggesting a higher percentage of false positives than false negatives.

| Model              | P            | R            | F1           |
|--------------------|--------------|--------------|--------------|
| ELMo-inaccurate    | 58.76        | 43.29        | 49.85        |
| ELMo-incomplete    | 71.76        | 78.03        | 74.77        |
| ELMo-hybrid        | 77.05        | 77.22        | 77.14        |
| BioELMo-inaccurate | 58.09        | 42.44        | 49.05        |
| BioELMo-incomplete | 73.30        | 78.19        | 75.67        |
| BioELMo-hybrid     | 74.74        | 79.69        | 77.14        |
| BERT-inaccurate    | 58.95        | 42.73        | 49.54        |
| BERT-incomplete    | 73.96        | <b>82.29</b> | 77.91        |
| BERT-hybrid        | 76.16        | 81.44        | 78.71        |
| BioBERT-inaccurate | 58.62        | 42.41        | 49.22        |
| BioBERT-incomplete | 76.98        | 80.66        | 78.77        |
| BioBERT-hybrid     | <b>77.23</b> | 81.53        | <b>79.32</b> |

Table 3: Evaluation of weak supervision methods; P: Precision, R: Recall, F1: F1 score.

| Span Type | Precision | Recall | F1    | # Occurrences |
|-----------|-----------|--------|-------|---------------|
| Reason    | 80.52     | 84.27  | 82.36 | 3459 (3687)   |
| Time      | 66.24     | 72.02  | 69.01 | 997 (1163)    |

Table 4: Performance of BioBERT-hybrid model and the number of occurrences of each span type in ground truths and predictions respectively.

## 6 Error Analysis

| Error Type                       | Reason | Time  |
|----------------------------------|--------|-------|
| Correct Label - Overlapping Span | 6.83   | 14.61 |
| Wrong Label - Correct Span       | 0.08   | 0.08  |
| Wrong Label - Overlapping Span   | 0.13   | 0.77  |
| Complete False Positive          | 13.77  | 23.12 |
| Complete False Negative          | 8.03   | 11.41 |

Table 5: Percentage of error types on the test set using the BioBERT-hybrid model.

To better understand the errors in predictions, we computed percentages of different types of errors (Table 5). The cases where the model predicted the right label but with an overlapping span (Correct Label-Overlapping Span) are mainly due to inconsistencies in annotations. The primary source of these inconsistencies is when annotators missed annotating a prepositional phrase or a time modifier phrase in the time span. Wrong label errors (Wrong Label - Correct Span, Overlapping Span) are minimal, suggesting that the model distinguishes between the time and reason spans very well.



Complete false positives and false negatives are the significant sources of errors for both reason and time spans and our qualitative analysis suggests that these cases often happen when multiple reason phrases and time phrases are present in the utterance window, but only a subset of them are valid. Because the task actually involves two different aspects, extracting reason and time mentions and spotting their confirmation clues, it may be difficult for the trained system to select exactly the confirmed reason or time mentions without explicitly modeling their relations. The ambiguity due to the oral nature of the conversations also makes it difficult to spot the confirmation clues.

Notably, we observe that the portion of complete false positives for the time span is significantly higher than reason spans. For example, the conversation in Figure 1 discusses several options for the appointment time, but the patient finally settles for Monday. The model often struggles with such cases and also extracts time mentions that are not confirmed. Using SpaCy’s NER, we find that 87% of these errors occurred when multiple time phrases are present, but not all are valid. The model may have difficulty with these cases because they amount to only 21.3% of the manually annotated time spans. Further, the annotated time spans are infrequent by a factor of three than the reason spans. These reasons explain why the F1 score on time span is significantly lower than the reason span.

## 7 Conclusion

In summary, we defined a novel task of extracting appointment spans from medical conversations, described our annotation methodology, and employed three weak supervision approaches to account for the limited set of annotations. Our proposed hybrid weak supervision approach showed improvement across all our experiments. Finally, our error analysis shows that a significant portion of the errors comes from false positives where the model has difficulty in identifying the correct span when multiple appointment reason or time mentions are present. In future work, we plan to study the data augmentation approaches as well as joint entity and relation extraction approaches to improve performance on difficult examples. We also plan to study the generalization of this work to automatic transcripts, whose transcription error rate may challenge entity detection.

## References

- B. G. Arndt, J. W. Beasley, M. D. Watkinson, J. L. Temte, W. J. Tuan, C. A. Sinsky, and V. J. Gilchrist. 2017. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *Ann Fam Med*, 15(5):419–426.
- Alan R Aronson. 2006. Metamap: Mapping text to the umls metathesaurus. *Bethesda, MD: NLM, NIH, DHHS*, 1:26.
- Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl\_1):D267–D270.
- Joanne L Callen, Johanna I Westbrook, Andrew Georgiou, and Julie Li. 2012. Failure to follow-up test results for ambulatory patients: a systematic review. *Journal of general internal medicine*, 27(10):1334–1348.
- A Chandio, Z Shaikh, K Chandio, SM Naqvi, and SA Naqvi. 2018. Can “no shows” to hospital appointment be avoided. *Clin Surg*, 3:1918.
- Angel X Chang and Christopher D Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. In *Lrec*, volume 3735, page 3740.
- Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. [One billion word benchmark for measuring progress in statistical language modeling](#). Technical report, Google.
- Chung-Cheng Chiu, Anshuman Tripathi, Katherine Chou, Chris Co, Navdeep Jaitly, Diana Jaunzeikare, Anjuli Kannan, Patrick Nguyen, Hasim Sak, Ananth Sankar, Justin Tansuwan, Nathan Wan, Yonghui Wu, and Xuedong Zhang. 2017. [Speech recognition for medical conversations](#). *CoRR*, abs/1711.07274.
- Roger Collier. 2017. Electronic health records contributing to physician burnout.
- Leila F Dantas, Julia L Fleck, Fernando L Cyrino Oliveira, and Silvio Hamacher. 2018. No-shows in appointment scheduling—a systematic literature review. *Health Policy*, 122(4):412–421.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- N Lance Downing, David W Bates, and Christopher A Longhurst. 2018. Physician burnout in the electronic health record era: are we ignoring the real cause?
- D. A. Drossman and J. Ruddy. 2019. Improving Patient-Provider Relationships to Improve Health Care. *Clin. Gastroenterol. Hepatol*.

- Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. 2019a. Extracting symptoms and their status from clinical conversations. *arXiv preprint arXiv:1906.02239*.
- Nan Du, Mingqiu Wang, Linh Tran, Gang Li, and Izhak Shafran. 2019b. Learning to infer entities, properties and their relations from clinical conversations. *arXiv preprint arXiv:1908.11536*.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Qiao Jin, Bhuwan Dhingra, William W Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. *arXiv preprint arXiv:1904.02181*.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. [BioBERT: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*, 36(4):1234–1240.
- Steven Y Lin, Tait D Shanafelt, and Steven M Asch. 2018. Reimagining clinical documentation with artificial intelligence. In *Mayo Clinic Proceedings*, volume 93, pages 563–565. Elsevier.
- Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.
- Martina Mookadam, Michael Grover, Chris Pullins, Mary Winscott, and Susan Pierce. 2016. Simple interventions improve the quality of a missed lab appointment process. *BMJ Open Quality*, 5(1).
- Dhruvesh Patel, Sandeep Konam, and Sai P. Selvaraj. 2020. [Weakly supervised medication regimen extraction from medical conversations](#).
- Noelle Junod Perron, Melissa Dominicé Dao, Nadia Comparini Righini, Jean-Paul Humair, Barbara Broers, Françoise Narring, Dagmar M Haller, and Jean-Michel Gaspoz. 2013. Text-messaging versus telephone reminders to reduce missed appointments in an academic primary care clinic: a randomized controlled trial. *BMC health services research*, 13(1):1–7.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Erik F Sang and Jorn Veenstra. 1999. Representing text chunks. *arXiv preprint cs/9907006*.
- Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.
- Benjamin Schloss and Sandeep Konam. 2020. Towards an automated soap note: Classifying utterances from medical conversations. In *Machine Learning for Healthcare Conference*, pages 610–631. PMLR.
- Sai P. Selvaraj and Sandeep Konam. 2019. [Medication regimen extraction from medical conversations](#).
- Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. [Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties](#). *Annals of Internal Medicine*, 165(11):753–760.
- Jannik Strötgen and Michael Gertz. 2010. Heidelberg: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324.
- S Ullah, S Rajan, T Liu, E Demagistris, R Jahrstorfer, S Anandan, C Gentile, and A Gil. 2018. Why do patients miss their appointments at primary care clinics. *J Fam Med Dis Pre*, 4:09.
- Zhi-Hua Zhou. 2018. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.