Alignement Annotation for Clinic Visit Dialogue to Clinical Note Sentence Language Generation

Wen-wai Yim†, Meliha Yetisgen*, Jenny Huang‡, Micah Grossman‡
† Augmedix, Inc., *University of Washington, ‡ University of California, Los Angeles
† 1161 Mission St 210, San Francisco 94103,
*850 Republican St, Seattle WA 98109,
‡ 405 Hilgard Avenue, Los Angeles, CA 90095
wenwai.yim@augmedix.com, melihay@uw.edu, jyh08@ucla.edu, micah.grossman@augmedix.com

Abstract

For every patient’s visit to a clinician, a clinical note is generated documenting their medical conversation, including complaints discussed, treatments, and medical plans. Despite advances in natural language processing, automating clinical note generation from a clinic visit conversation is a largely unexplored area of research. Due to the idiosyncrasies of the task, traditional methods of corpus creation are not effective enough approaches for this problem. In this paper, we present an annotation methodology that is content- and technique- agnostic while associating note sentences to sets of dialogue sentences. The sets can further be grouped with higher order tags to mark sets with related information. This direct linkage from input to output decouples the annotation from specific language understanding or generation strategies. Here we provide data statistics and qualitative analysis describing the unique annotation challenges. Given enough annotated data, such a resource would support multiple modeling methods including information extraction with template language generation, information retrieval type language generation, or sequence to sequence modeling.

Keywords: Corpus, Dialogue

1. Introduction

Two trends drive significant interest in automating the process of medical documentation: a growing shortage of clinicians in the United States and a rise in clinician burnout rates due to health information technology-related stress (Gidwani et al., 2017; AAMC, 2019; Gardner et al., 2019). Clinicians today are responsible for more than just the well-being of their patients. They must also complete documentation on their time present with the patient, diagnoses made, and treatments prescribed in the electronic medical record. Medical scribes, who can assist clinicians with completing medical documentation, are one way clinicians can unburden themselves from documentation responsibilities. But the cost of employing a medical scribe, estimated between $49K (onsite) and $23K (virtual) a year (Brady and Shariff, 2013), is prohibitive.

Despite recent natural language processing (NLP) advancements, such as improved speech to text, deep learning NLP modeling, and greater availability of clinical NLP resources, the task of converting a clinic visit conversation into its corresponding clinical note remains challenging. True comprehension of the clinical situation discussed in a visit requires many difficult aspects of language understanding and generation, such as summarizing over multiple statements and question-answer responses. Clinical note generation also depends on input from outside factors, e.g. electronic medical record data, templates, and patients’ reported visit complaints. When a scribe is present, intake information can come from both a clinic visit conversation dialogue as well as direct communication with a medical scribe. This fluidity in dialogue and sourcing adds more complexity to the problem. Parts of the conversation may involve comforting a patient, clarifying information, or extracting information. What ultimately becomes documentation-worthy is often highly specialty-, institution-, and provider-specific. Properly trained NLP models will require large annotated corpora to achieve this ambitious goal and capture the ways non-sequitur statements or question-answer summarizations inform the final clinical note content. Previous annotation methodologies do not address the unique nature of this problem. Moreover, due to concerns over patient privacy, medical conversation data and clinical note data, each by themselves, are low-resource domains.

In this paper, we introduce a novel annotation methodology which matches clinical note sentences to grouped sets of dialogue sentences. The goal is to build a corpus of annotated data which can support a variety of techniques, including information extraction with template language generation or sequence to sequence methods. With such a corpus one could build systems to generate clinical notes from clinic visit audio automatically. This annotation approach flexibly adapts to the significant variability between provider, specialty, and institution. To our knowledge, this is the first work to attempt such a content- and technique-agnostic systematic annotation methodology for this task.

2. Background

While unique institutions and departments may have different practices, a pattern of common steps emerges in note creation. Pre-charting is the first step, done prior to the start of a visit: an appropriate note template is selected and populated where necessary with pertinent information related to the patient as well as reason for visit, indicated from scheduling. During the actual visit, the clinician converses with a patient to gather details of the problem, diagnose illnesses, and discusses treatments and plans. This is the data capture step. At times the clinician will prompt for medical context from the patient. At other times, the clinician col-
Table 1: Implied visit state and note section output dependencies

- **Discovering reason for visit** / **Verbal examination**
- **History of Present Illness (HPI)**
- **Review of System (ROS)**
- **Social History (SHs)**
- **Routine Health Maintenance (RHM)**
- **Physical examination**
- **Physical Examination**
- **Detailing treatment or further investigation**
- **Assessment and Plan**
- **Impression**

Table 1 shows example visit states and note section destinations. Table 2 shows the resulting note sentence from portions of a dialogue. Not only must information be extracted across two question answer adjacency pairs (and 4 turns), but the second question answer pair does not give an explicitly mentioned subject. The nature of conversations further complicates the matter, as multiple threads can happen at the same time covering different areas of the visit state. An example of this is shown in Table 3 during a physical examination, the clinician and patient may discuss other matters.

A full abbreviated dialogue exchange and its resulting clinical note are shown in Table 4 to illustrate such complexities. The note is organized into sections. Several non-continuous parts of the dialogue may interweave (lines 84-

---

Table 2: Summarization into note sentence across four turns of questions and answers

| Note                                                                 | Dialogue                                                                 |
|----------------------------------------------------------------------|-------------------------------------------------------------------------|
| She declines the pneumonia vaccine.                                  | 28 | Doctor: Have you had a pneumonia vaccine? |
|                                                                      | 29 | Patient: No, I don’t think so.          |
|                                                                      | 30 | Doctor: Alright, do you want one?       |
|                                                                      | 31 | Patient: No.                            |

Table 3: Intertwining visit states and anaphora

- **Discovering reason for visit** / **Verbal examination**
- **History of Present Illness (HPI)**
- **Review of System (ROS)**
- **Social History (SHs)**
- **Routine Health Maintenance (RHM)**
- **Physical examination**
- **Physical Examination**
- **Detailing treatment or further investigation**
- **Assessment and Plan**
- **Impression**

Table 4: Example annotations (right) for corresponding clinical note (left) and dialogue (middle). The same colors indicate matched associations.
Anaphora is ubiquitous (line 89). The order of appearances in note and dialogue often don’t correspond, leading to many crossing annotations.

3. Related Work

While some work exists on doctor-patient conversation analysis (Byrne and Long, 1977; Rainbault et al., 1975; Drass, 1982; Cerny, 2007), annotation (Wang et al., 2018), and dialogue topic classification (Rajkomar et al., 2019), few explore the relationship between a patient visit’s dialogue and a clinical note. We describe three groups with some coverage of the problem.

In (Kazi and Kahanda, 2019), the authors studied generating psychiatric case note sentences from doctor-patient conversation transcripts. Their work classified doctor-patient response pairs into semantic categories (e.g. client details, family history), then used a rule-based language processing system/tool/etc. to paraphrase the text into formal clinical text. Though an interesting idea, the data set was small (18 transcripts) and the authors do not assess the performance of their natural language generation.

In (Jeblee et al., 2019), the authors use 800 conversations to perform several classifications including: utterance type (e.g. question, statement), temporal and clinical entity extraction, attribute classification, classification of entries to a SOAP format and classification of primary diagnosis. Sentence generation was left for future work. This approach makes strong technique commitments, assumes a fixed clinical note template output, and does not lend well to support paraphrasing techniques.

In (Finley et al., 2018a), members of the EMR.AI team describe one intended approach to the problem by bridging information from clinic visit dialogue, first by classifying conversation sentences by intended note sections, then applying information extraction techniques. This data is then used to fill note templates generated by finite-state grammars. In another work (Finley et al., 2018b), they describe their method to automatically produce a parallel machine translation corpus for the special case of dictations to clinical note letter, but this focuses on just a narrow portion of the general problem.

We posit that the task of clinical note generation based on dialogue is best represented as an amalgamation of different language transformations and thus the annotation efforts should not be tied to specific end to end methods. Our proposed method associates a note sentence to associated dialogue sentence sets using different tags (e.g. DICTATION, QA, STATEMENT, etc) and provides a higher level ordering of these sets. Compared to (Kazi and Kahanda, 2019), (Jeblee et al., 2019), and (Finley et al., 2018a), we actually annotate the final note content output. Compared to the work in (Finley et al., 2018b), we manually create our alignments and do it for the entire conversation and note. Therefore our dataset does not rely on a specific sequence of domain-dependent NLU tasks or a specific relationship between the extracted information and the final output (e.g. template-filling), nor assumes a narrow part of the problem.

4. Data

The data comes from 66 mock patient visit materials used for training scribes. Though simulated, they were created to mimic real interactions. Material for each visit includes an audio recording and an associated clinical note. Audio recordings, 9.5 minutes duration on average with minimum, maximum and 50th-percentile at 2.1, 19.0, 9.5 minutes respectively, were run through Azure’s speech to text service (azure.microsoft.com/en-us/services/cognitive-services/speech-to-text, 2019). The transcript output was a segmented and corrected for word errors. The dialogue transcript was sentence tokenized based on punctuation. Notes were tokenized using spaCy (spacy.io, 2019). The total number of dialogue sentences was 11465, with a vocabulary of 4706 words, with averages of 95 number of turns and 174 sentences per visit. The primary speakers were the clinician-primary, with 7133 sentences, 3135 turns; the patient, 4050 sentences, 2937 turns; in addition to several other speakers (Table 5). Clinical notes had a total of 3181 sentences, a mean of 48 sentences per visit, and a vocabulary of 2873 words.

| Speaker            | Sentences | Turns |
|--------------------|-----------|-------|
| clinician_primary  | 7133      | 3135  |
| patient            | 4050      | 2937  |
| guest_patientfamily| 242       | 155   |
| other              | 40        | 27    |

Table 5: Dialogue speaker breakdown

5. Annotation Methodology

The annotation methodology was created iteratively by a diverse group of clinical NLP experts, medical scribes, and

---

1. Anaphora is the phenomenon when an expression can only be understood within context of another expression

2. We define turn as a unit of continuous spoken language by one speaker before a speaker transition
a trained computer scientist with healthcare experience. The inter-annotator agreement was performed by the latter three. The corpus was annotated using an in-house software shown in Figure 2.

During annotation, each sentence of the clinical note can be attributed to multiple sets of sentences from the dialogue. Each labeled association of this type creates a set (In the software, to accomplish this, an annotator can select one note sentence and many dialogue lines and click a tag, e.g. STATEMENT, to create a new set which appears in the top middle panel). Each set should only include consecutive sentences of the same label, of the same speaker. The exception to this is the QA label which is expected to be 1 question, 1 answer, not necessarily consecutive. The QA label may also be used to mark a question without an answer. High level tags can only act upon other sets of the same label, of the same speaker. The exception to this is the QA label which is expected to be 1 question, 1 answer, not necessarily consecutive.

5.1. Annotations

Tags Addressing the Scribe
When the clinician is speaking outside of a conversation mode, we employ multiple tags to capture possible fundamental linguistic differences, such as when performing dictation or speaking to a scribe.

- **COMMAND**: Commands that specify document changes, which are attached to note section headers, e.g. “Use normal PE”.
- **DICTATION**: Spoken information intended to be dictated, e.g. “Mildly elevated liver enzymes, likely related to increased alcohol use the week prior to the labs”.
- **STATEMENT2SCRIBE**: Other spoken information to scribe.

Template-related Tags
Clinical notes are often built on prefabricated templates. An annotator is assumed to have access to the originating note template and its default values. In cases when the note sentence comes from the template’s default values, an annotator will apply one of the tags below.

- **INFERRER-DIALOGUE**: Non-explicit spoken information that can imply a template default (e.g. if in the dialogue it is known that a system is checked “let me listen to your lungs”) but there is no explicit mention of abnormality, we can infer that the default value is correct “lungs: clear to auscultation bilaterally”.
- **INFERRER-OUTSIDE**: Otherwise marks note sentences as coming from a template default.

Conversation Tags
During conversation, we identify two major modes of information exchange: question-answer and statements.

- **STATEMENT**: Statements spoken during clinic visit conversation, e.g. “She is here for her annual physical”.
- **QA**: Question-answer modes during clinic visit conversation, e.g. “Any vomiting or diarrhea? No.”.

Higher-level Tags
To mark some small amount of structure between normal sets, we use several higher level tags.

- **GROUP**: Used to group together discontinuous sets that are anaphoric.
- **REPEATS**: Used to indicate when a note sentence can be separately derived from different dialogue sets.

A clinical note’s sentence can be left unattributed when its content is only derivable from outside knowledge, e.g. laboratory data. An already labeled note sentence may additionally be marked with an INCOMPLETE label when some information is unidentifiable from the dialogue. To make the annotation task tractable, we implemented sev-

---

3This is consistent with, though a simplification of Todd’s classification of doctor-patient communication speech acts: statements, questions, answers, directives, and reactives; for which we are not interested in the latter two (Cerny, 2007).
Table 6: Salience rule example. Only the first sentence, which additionally indicates “last night”, should be associated in this case.

Table 7: Repeat examples

5.2. Inter-annotator Agreement

A single match between one note sentence and its associated dialogue sentences can be represented as a match tree, as shown in Figure 5. We evaluate matches according to three metrics: unlabeled triple, path, and span metrics. The first is a simple unlabeled f1 metric of selected dialogue sentences. An example from Figure 5 is ‘mock_patient_01\note_18\[86,87,88\]’ (Sampson and Babarczy, 2003). These reflect measures for simple matches as well as vertical and horizontal evaluations.

6. Quantitative Analysis

Table 8 shows the final inter-annotator agreement for a total of 10 clinic visits. Unsurprisingly, unlabeled triple had the highest agreement, with lower agreement for more complex metrics. Table 9 shows unlabeled triple agreements broken down by category. INFERRED-DIALOGUE tagged sentences were low as it requires judgement over what dialogue can infer the information from the note. We purposefully did not specify very specific guidelines for it, as both INFERRED and INFERRED-DIALOGUE are chiefly to mark parts of the note that came from templates and its default values.

Out of all available note sentences, 81 ± 1 % were marked. In contrast, out of all dialogue lines 39 ± 11 % were marked. To generate clinical notes from dialogue, the note’s author has to filter or aggregate significant amounts of information. Max, min, median tree heights were 6, 4, 3 respectively. This supports our annotation design which encourages shallower trees, by only annotating the most salient evidence. For GROUP sets, 68% include 2 sets, 22%, 6%, and 2% for groups of 3, 4, and 5 sets respectively. For REPEAT sets, 92% contain two sets, 8% three sets. The percentage of note and dialogue sentences with at least one label is shown in Table 10. In contrast, Table 11 shows the frequencies of a note sentences with the total number of associated tags. While note sentences with only STATEMENT dialogue information occurred the most frequently, this made up only 16% of all sentences, while DICTION only 3%. Interestingly as high as 21% of note sentences required a GROUP label and that composition of multi-tagged note sentences were the most frequent. Together, this suggests that summarizing over multiple types of dialogue source information (e.g. STATEMENT and QA) is often required for note content generation.

We also measured the difference between the maximum and minimum dialogue lines for the top 3 frequent labels,
crosses with at least three other note sentences. Since the preamble and after-visit clarifications may often provide salient information, if we did not count COMMAND, DICTATION, or STATEMENT2SCRIBE (the NON-SCRIBE) column, we would still find 60% of note sentences have information that crosses with at least three other note sentences’ annotations. In all, the high percentages show that cross matches occurs frequently which would make automatic sentence alignment challenging.

Table 10: Percentage of sentences per label

| Label-set          | Freq | %   | Cum. % |
|--------------------|------|-----|--------|
| [STATEMENT]        | 496  | 16  | 16     |
| [INFERRED-OUTSIDE]| 395  | 12  | 28     |
| [QA]               | 305  | 10  | 38     |
| [QA,STATEMENT,GROUP]| 279 | 9   | 46     |
| [STATEMENT2SCRIBE]| 238  | 7   | 54     |
| [STATEMENT,GROUP]  | 189  | 6   | 60     |
| [INFERRED-DIALOGUE]| 106  | 3   | 63     |
| (DICTATION)        | 98   | 3   | 66     |

Table 11: Top 8 occurring tagsets per note sentence. Column 4 is the sum percentage top to down.

which is informative for understanding how to group dialogue (Table 12). Typically, most occurrences of STATEMENT sets were single line (n=0), while the most frequent of QA was over two lines (n=1); though sometimes a question appeared without an answer (n=0). This shows that while most paired associations are a single sentence, there is spread of information across proximal sentences in a dialogue for STATEMENT. For QA, while many times the answer to a question can be found after the question, this is not always the case. Finally, the spread of required dialogue lines for GROUP label sentences (which account for 21% of sentences) suggests that to capture all related information per note sentence requires gathering related sentences spread across the dialogue.

Table 12: Dialogue ranges (n lines) for associated top occurring labels per note sentence.

| n   | STATEMENT | QA | GROUP |
|-----|-----------|----|-------|
| 0   | 1,546     | 73 | 0     |
| 1   | 253       | 110| 20    |
| 2   | 66        | 131| 109   |
| 3   | 20        | 38 | 104   |
| 4   | 7         | 10 | 81    |
| ≥5  | 7         | 8  | 202   |

Analyzing the similarity of matched text between dialogue and note, we calculate the jaccard coefficient for unigrams and UMLS concept identifiers, tagged with Metamap (Aronson and Lang, 2010), for the associated texts. The low similarity scores shown in (Table 13a) suggests that to get full matching context, simple similarity algorithms would be challenging. To quantify alignment difficulty, we calculate the percentages of sentences that cross n other sentences for tags exclusively not directed at the scribe, as well as for all sentences (Table 13b). For example, for n=3, we see that 76% of note sentences have evidence that

Table 13a: Data statistics of paired associations (a) Jaccard similarity

Table 13b: Data statistics of paired associations (b) % crossing annotations

7. Qualitative Analysis

In this section, we give qualitative descriptions along with examples to give the reader further insight into annotation disagreements. Below we describe several features of the annotation problem that present annotation challenges: (1) tag ambiguities related to the domain, (2) the problem of aggregating information and annotating over dialogues, and (3) the difficulty of using the GROUP label in the face of anaphoric references.

7.1. Domain-related Tag Ambiguities

On manual analysis, there were some common domain-related tag disagreement issues: (1) confusion between STATEMENT2SCRIBE vs DICTATION, (2) STATEMENT2SCRIBE vs STATEMENT, and (3) INFERRED-DIALOGUE and INFERRED-OUTSIDE. For the most part when it is clear the clinician is speaking to the scribe and word-for-word translation is required, it is very apparent that associated label should be a DICTATION. However, there are some cases when there is a blend of speech for which the spoken information should be paraphrased and some in which direct copy would work. Though this is not a large problem (the agreement for DIC-TATION is amongst the highest) this is an interesting phenomenon. An example of this is shown in Table 14. Confusion between STATEMENT2SCRIBE and STATEMENT occurs for cases in which a clinician can be interpreted as either speaking to the scribe or to the patient. This is more typical during the physical exam.

Table 14: Confusable STATEMENT2SCRIBE and DICTATION examples
The INFERRED-* tags denote parts of the clinical note that belong to default values of templates which is never explicitly mentioned in the dialogue. INFERRED-DIALOGUE is meant to capture parts of a dialogue that may suggest certain information. However, what is considered to be suggestive is somewhat subjective at times.

### 7.2. Annotating Over Lengthy Dialogue

A hallmark of our task is the requirement of the annotator to identify the most representative information of a clinical note sentence across an entire dialogue. This is especially difficult given the length of dialogues (on average 174 sentences per visit). This problem is compounded by the requirements of identifying the most salient information across many repeats of the same topics with various levels of information completion, as well as the unpredictable ordering of topics in the dialogue as compared to their appearance in the clinical note.

One alternative would include having annotators mark only first appearance. Another strategy would be to require annotators to mark every relevant sentence. In the former case, this strategy would forgo easy capture of variations within the same conversation – it would also forgo future abilities to automatically measure differences in expression of the same information within the same conversation. In the latter case, the amount of required annotations would vastly increase; furthermore the paired associated dialogue text would then contain much more repeating bits of information which would make the paired association less readily useful for artificial learning applications.

### 7.3. Grouping and Anaphora

Anaphora is ubiquitous in natural language, and especially in dialogue. The GROUP tag is used to mark such instances. For the special case in which referring expressions for pronouns and determiners need to be captured, the annotator is required to additionally mark the closest dialogue passage with the subject and connect the two sets with a GROUP tag. At times, this may be far from the actual information and may be ambiguous.

An example is shown in Table [15]. Different annotators may identify different passages to what constitutes an acceptable referential named entity. For cases in which additional anaphoric connections are required on top of the already identified referring expressions for pronouns and determiners, annotators – according to guidelines – are meant to connect to higher nodes. The exact hierarchy ordering can be easily perturbed amongst different annotators. Table [16] shows a complex group example.

### 8. Discussion

Given the complexity and innate ambiguity of the task, we believe our agreements are good. Inconsistency between annotator’s associations does not signify incorrectness (e.g. a sentence can have equally correct constituency parses). Table [17] shows three annotator markings for the same sentence in which all have annotation errors. The resulting annotation match trees are very different, but the content from each annotator markings are accurate.

| Note | Dialogue |
|------|----------|
| She reports that her discharge has been worsening over the past two days. | setid[2] - STATEMENT |
| 2 | patient: Yes I have some discharge coming from my left eye. | setid[2] - STATEMENT |
| 9 | patient: Well it’s been feeling funny the past week waking up, but in the past couple of days I couldn’t even open it when I woke up. |

**Table 15: GROUP label used for pronoun and determiners**

| Note | Dialogue |
|------|----------|
| Could be viral infection or a side effect of cephalaxin, but also concerned regarding C. difficile infection. | setid[59] - STATEMENT |
| 59 | clinician:primary: See you know any antibiotics can give you a little diarrhea. |

**Table 16: Complex group example**

For example, A1/A2 triple, path, and span F1 agreements for these instances are 0.50, 0.13, 0.30; A1/A3, 0.77, 0.31, 0.54; A2/A3, 0.15, 0.00, 0.07. In comparison, the work by (Hwang et al., 2015) marking alignments between Simple Wikipedia and Wikipedia pages, which we believe to be closest to our task, achieved an annotator agreement of 0.68 Kappa for 46 articles and 67,853 sentence pairs. Our analogous agreement metric of simple match f1 was on average 0.73, which is consistent with this baseline.

In future iterations, we will add user-friendly improvements, including text search and highlights of clinically important elements to better assist annotating. Though sentences are the main unit of alignment here, this may be practically adjusted for real data: e.g. an automatically generated table in the note should be considered one unit instead of trying to divide the table into sentences, etc. While the data and analysis here was created from mock patient visits, the dialogue content should approximate actual real patient visits.
### 9. Conclusions

In this work, we introduce a new annotation methodology for marking paired associations between a clinic visit dialogue and its complete clinical note. Given a large enough corpus of annotated data, the noise related to the difficulty of the task can be overcome to support multiple methods such as information extraction with templated generation, information retrieval type language generation, or sequence to sequence modeling. The same corpus can be used to train models for reordering generated sentences into proper required underlying clinical note structure or for building classifiers for slots in predetermined clinical note templates. In future work, we will apply this annotation approach for a large corpus of real patient visits to aid in clinical note content creation. Furthermore, we will train matching algorithms to assist annotation. Finally, after incorporating clinical note generated suggestions to aid our scribing operations, we can additionally use user-feedback for re-ranking output.

### 10. Acknowledgments

We would like to thank the Augmedix Scribe Operations who was responsible for source material creation and cleaning. We especially want to thank Jane Kwong for early contributions to the annotation schema creation and iteration.
11. Bibliographical References

AAMC. (2019). New findings confirm predictions on physician shortage.

Aronson, A. R. and Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association: JAMIA*, 17(3):229–236.

azure.microsoft.com/en-us/services/cognitive-services/speech-to-text. (2019). Speech to text API | microsoft azure.

Brady, K. and Shariff, A. (2013). Virtual medical scribes: making electronic medical records work for you. *J Med Pract Manage*, 29(2):133–6.

Byrne, J. F. and Long, P. (1977). Doctors talking to patients. *Psychological Medicine*, 7(4):735.

Cerny, M. (2007). On the function of speech acts in doctor-patient communication. *Linguistica*.

Drass, K. A. (1982). Negotiation and the structure of discourse in medical consultation. *Sociology of health & illness*.

Finley, G., Edwards, E., Robinson, A., Brenndoerfer, M., Sadoughi, N., Fone, J., Axtmann, N., Miller, M., and Suendermann-Oeft, D. (2018a). An automated medical scribe for documenting clinical encounters. In *NAACL-HLT*.

Finley, G., Salloum, W., Sadoughi, N., Edwards, E., Robinson, A., Axtmann, N., Brenndoerfer, M., Miller, M., and Suendermann-Oeft, D. (2018b). From dictations to clinical reports using machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 121–128, New Orleans - Louisiana, June. Association for Computational Linguistics.

Gardner, R., Cooper, E., Haskell, J., Harris, D., Poplau, S., Kroth, P., and Linzer, M. (2019). Physician stress and burnout: the impact of health information technology. *Jamia*, 26:106–114.

Gidwani, R., Nguyen, C., Kofoed, A., Carragee, C., Rydel, T., Nelligan, L., Sattler, A., Mahoney, M., and Lin, S. (2017). Impact of scribes on physician satisfaction, patient satisfaction, and charting efficiency: A randomized controlled trial. *Annals of Family Medicine*, 15(5):427–433.

Hwang, W., Hajishirzi, H., Ostendorf, M., and Wu, W. (2015). Aligning sentences from standard wikipedia to simple wikipedia. In *HLT-NAACL*.

Jeblee, S., Khan Khattak, F., Crampton, N., Mamdani, M., and Rudzicz, F. (2019). Extracting relevant information from physician-patient dialogues for automated clinical note taking. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 65–74. Association for Computational Linguistics.

Kazi, N. and Kahanda, I. (2019). Automatically generating psychiatric case notes from digital transcripts of doctor-patient conversations. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 140–148, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation.

Raimbault, G., Cachin, O., Limal, J. M., Eliacheff, C., and Rappaport, R. (1975). Aspects of communication between patients and doctors: an analysis of the discourse in medical interviews.

Rajkomar, A., Kannan, A., Chen, K., Vardoulakis, L., Chou, K., Cui, C., and Dean, J. (2019). Automatically charting symptoms from patient-physician conversations using machine learning. 179(6):836–838.

Sampson, G. and Babarczy, A. (2003). A test of the leaf-ancestor metric for parse accuracy | natural language engineering | cambridge core. 9(4):365–380.

spacy.io. (2019). spaCy Â· industrial-strength natural language processing in python.

Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., and Yi, L. (2014). UM-corpus: A large english-chinese parallel corpus for statistical machine translation. In *LREC*.

Tiedemann, J. (2011). Bitext alignment. *Synthesis Lectures on Human Language Technologies*, 4(2):1–165.

Wang, N., Song, Y., and Xia, F. (2018). Constructing a Chinese medical conversation corpus annotated with conversational structures and actions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).