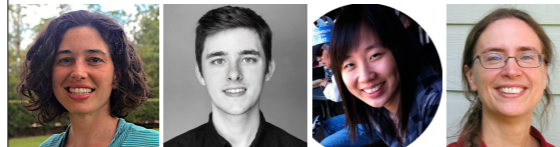


# From News to Medical: Cross-domain Discourse Segmentation

Elisa Ferracane<sup>1</sup>, Titan Page<sup>2</sup>, Jessy Li<sup>1</sup>, Katrin Erk<sup>1</sup>

<sup>1</sup>University of Texas at Austin

<sup>2</sup>University of Colorado Boulder



I'm presenting joint work with Titan Page at UC Boulder, and Jessy Li and Katrin Erk at UT Austin. Today I'll be talking about what happens when you use a news-trained discourse segmenter on medical data. And to be explicit, when I say discourse segmentation, I'm referring to Rhetorical Structure Theory or RST segmentation.

# Rhetorical Structure Theory (RST)

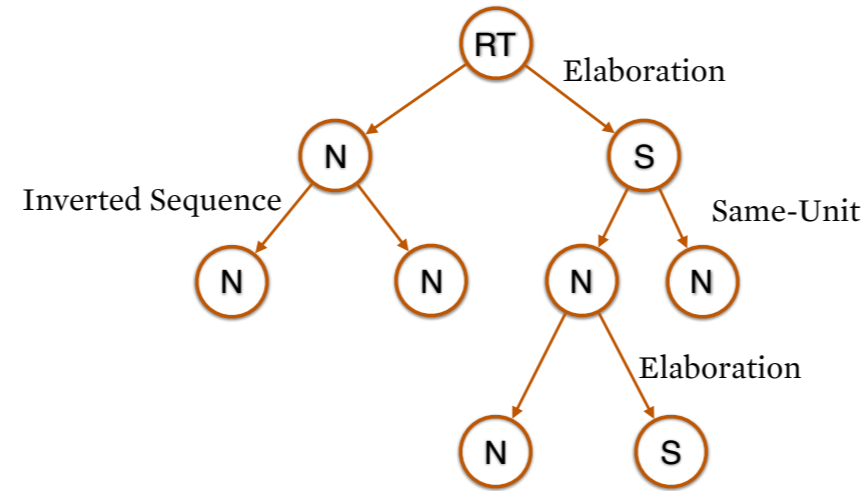
In RST, you convey the rhetorical organization of a document with a labelled tree structure. Here is an excerpt from a WSJ news article.

## Rhetorical Structure Theory (RST)

*Three new issues begin trading on the New York Stock Exchange today, and one began trading on the Nasdaq/National Market System last week. On the Big Board, Crawford & Co., Atlanta, (CFD) begins trading today.*

In RST, you convey the rhetorical organization of a document with a labelled tree structure. Here is an excerpt from a WSJ news article.

# Rhetorical Structure Theory (RST)



And here is the RST tree you would get. But the first part in creating this tree is to segment the document into elementary discourse units or EDUs.

# Task: Discourse segmentation

And our work focuses on this task. Let's say I ask you to segment this document and give you a few rules of thumb for what an EDU is: roughly it's a clause or a parenthetical.

## Task: Discourse segmentation

*Three new issues begin trading on the New York Stock Exchange today, and one began trading on the Nasdaq/National Market System last week. On the Big Board, Crawford & Co., Atlanta, (CFD) begins trading today.*

And our works focuses on this task. Let's say I ask you to segment this document and give you a few rules of thumb for what an EDU is: roughly it's a clause or a parenthetical.

## Task: Discourse segmentation

*[Three new issues begin trading on the New York Stock Exchange today,][and one began trading on the Nasdaq/National Market System last week.][On the Big Board, Crawford & Co., Atlanta,][((CFD))][begins trading today.]*

Using these heuristics, let's segment our document. Lo and behold, we exactly match the gold boundaries. This is an easy task!

## Task: Discourse segmenta



*[Three new issues begin trading on the New York Stock Exchange today,] [and one began trading on the Nasdaq/National Market System last week.] [On the Big Board, Crawford & Co., Atlanta,] [(CFD)] [begins trading today.]*

Using these heuristics, let's segment our document. Lo and behold, we exactly match the gold boundaries. This is an easy task!



# Task: Discourse segmentation

In fact, discourse segmentation is usually treated as a solved task. Most RST parsers evaluate only on gold EDUs and don't even bother including an automated segmenter.

## Task: Discourse segmentation

- Usually treated as a ***solved*** task (F1=94.3)

In fact, discourse segmentation is usually treated as a solved task. Most RST parsers evaluate only on gold EDUs and don't even bother including an automated segmenter.

## Task: Discourse segmentation

- Usually treated as a **solved** task (F1=94.3)
- Many RST parsers:

In fact, discourse segmentation is usually treated as a solved task. Most RST parsers evaluate only on gold EDUs and don't even bother including an automated segmenter.

## Task: Discourse segmentation

- Usually treated as a **solved** task (F1=94.3)
- Many RST parsers:
  - evaluate **only on gold** segmented data
  - include **no automated segmenter**

In fact, discourse segmentation is usually treated as a solved task. Most RST parsers evaluate only on gold EDUs and don't even bother including an automated segmenter.

# Task: Discourse segmentation

However, using automatically segmented instead of gold EDUs does degrade results by 10% on the the downstream tasks for creating the rest of the tree structure—span, nuclearity and relation labeling. Furthermore, the few automated segmenters that are available are all trained on only one domain: news.

## Task: Discourse segmentation

- but if using automatically segmented vs. gold EDUs, results **degrade by 10%** on span, nuclearity, relation labeling tasks [Feng, 2015]

However, using automatically segmented instead of gold EDUs does degrade results by 10% on the the downstream tasks for creating the rest of the tree structure—span, nuclearity and relation labeling. Furthermore, the few automated segmenters that are available are all trained on only one domain: news.

## Task: Discourse segmentation

- but if using automatically segmented vs. gold EDUs, results **degrade by 10%** on span, nuclearity, relation labeling tasks [Feng, 2015]
- all automated segmenters are **trained on news**

However, using automatically segmented instead of gold EDUs does degrade results by 10% on the the downstream tasks for creating the rest of the tree structure—span, nuclearity and relation labeling. Furthermore, the few automated segmenters that are available are all trained on only one domain: news.

## Task: Discourse segmentation



- but if using automatically segmented vs. gold EDUs, results **degrade by 10%** on span, nuclearity, relation labeling [Feng, 2015]
- all automated segmenters are **trained on news**

However, using automatically segmented instead of gold EDUs does degrade results by 10% on the the downstream tasks for creating the rest of the tree structure—span, nuclearity and relation labeling. Furthermore, the few automated segmenters that are available are all trained on only one domain: news.



# Task: Discourse segmentation for medical domain

So what happens if I want to segment EDUs on a domain that isn't news? Let's take the medical domain. We focus on medical because it has already sparked a strong interest in the discourse research community (take, for example the Biomedical Discourse Relation Bank for Penn Discourse Treebank-style parsing) and because it wide applications in the real world. Our first research question is to understand the difficulties that news-trained segmenters have on medical data. Next, we look at 3 different segmenters with different features. How do the features of the segmenter impact the type of errors we get on medical? Third, we want to understand patterns in inter-annotator agreement and relate those to the performance of the segmenter in each of the medical data sections. To answer all these questions, we naturally need a corpus of segmented medical data.

## Task: Discourse segmentation for medical domain

1) What are difficulties of *news-trained* segmenters on *medical* data?

So what happens if I want to segment EDUs on a domain that isn't news? Let's take the medical domain. We focus on medical because it has already sparked a strong interest in the discourse research community (take, for example the Biomedical Discourse Relation Bank for Penn Discourse Treebank-style parsing) and because it wide applications in the real world. Our first research question is to understand the difficulties that news-trained segmenters have on medical data. Next, we look at 3 different segmenters with different features. How do the features of the segmenter impact the type of errors we get on medical? Third, we want to understand patterns in inter-annotator agreement and relate those to the performance of the segmenter in each of the medical data sections. To answer all these questions, we naturally need a corpus of segmented medical data.

## Task: Discourse segmentation for medical domain

- 1) What are difficulties of *news-trained* segmenters on *medical* data?
- 2) How do *features* of the segmenter impact the *type of errors* seen in medical?

So what happens if I want to segment EDUs on a domain that isn't news? Let's take the medical domain. We focus on medical because it has already sparked a strong interest in the discourse research community (take, for example the Biomedical Discourse Relation Bank for Penn Discourse Treebank-style parsing) and because it wide applications in the real world. Our first research question is to understand the difficulties that news-trained segmenters have on medical data. Next, we look at 3 different segmenters with different features. How do the features of the segmenter impact the type of errors we get on medical? Third, we want to understand patterns in inter-annotator agreement and relate those to the performance of the segmenter in each of the medical data sections. To answer all these questions, we naturally need a corpus of segmented medical data.

## Task: Discourse segmentation for medical domain

- 1) What are difficulties of *news-trained* segmenters on *medical* data?
- 2) How do *features* of the segmenter impact the *type of errors* seen in medical?
- 3) What is the relationship between *annotator agreement* and *segmenter performance* for different types of medical data?

So what happens if I want to segment EDUs on a domain that isn't news? Let's take the medical domain. We focus on medical because it has already sparked a strong interest in the discourse research community (take, for example the Biomedical Discourse Relation Bank for Penn Discourse Treebank-style parsing) and because it wide applications in the real world. Our first research question is to understand the difficulties that news-trained segmenters have on medical data. Next, we look at 3 different segmenters with different features. How do the features of the segmenter impact the type of errors we get on medical? Third, we want to understand patterns in inter-annotator agreement and relate those to the performance of the segmenter in each of the medical data sections. To answer all these questions, we naturally need a corpus of segmented medical data.

# Medical Corpus

Because no such corpus exists in English. We take 2 clinical trial reports and divide them into their respective sections (like Introduction, Results) for a total of 11 documents. Myself and my co-author Titan trained on the RST-DT and followed their same guidelines to annotate this corpus and we achieved a high inter-annotator agreement.

# Medical Corpus

- First RST-segmented medical corpus in English
- 11 documents: divided 2 clinical trial reports from PubMed Central into sections
- annotated by two trained Linguists ( $\kappa=0.90$ )

Because no such corpus exists in English. We take 2 clinical trial reports and divide them into their respective sections (like Introduction, Results) for a total of 11 documents. Myself and my co-author Titan trained on the RST-DT and followed their same guidelines to annotate this corpus and we achieved a high inter-annotator agreement.

## Introduction

[Patients with depression often suffer medical and psychiatric comorbidity.][{ 1,2 }]

Here are some excerpts from the corpus. In the introduction, we'll see a straightforward and simple sentence, but also an artifact that never comes up in news: citations. We treat these as parenthetical and always segment them. The next excerpt is from the discussion section. Immediately you can see a more complex discourse with 6 edus in this one sentence. We'll touch more on these differences later when we analyze automated and human errors broken down by section.

## Introduction

[Patients with depression often suffer medical and psychiatric comorbidity.][{ 1,2 }]

## Discussion

[In view of the lack of effectiveness][shown in our study][and the costs][associated with provision of a non-monetary incentive,][investigators should be deterred from utilising this particular strategy][to improve response rates from medical practitioners.]

Here are some excerpts from the corpus. In the introduction, we'll see a straightforward and simple sentence, but also an artifact that never comes up in news: citations. We treat these as parenthetical and always segment them. The next excerpt is from the discussion section. Immediately you can see a more complex discourse with 6 edus in this one sentence. We'll touch more on these differences later when we analyze automated and human errors broken down by section.



# Medical Corpus

Corpus	Docs	Tokens	Sents	EDUs
Medical	11	3356	169	399

To summarize, these are the stats on the small medical corpus we annotated. In order to compare to news

## Medical Corpus

Corpus	Docs	Tokens	Sents	EDUs
Medical	11	3356	169	399
RST-DT Small	11	4009	159	403

We sampled the same number of similarly sized documents from the RST DT and you can see the stats are very similar in terms of tokens, sentences and EDUs.

# Task: Discourse segmentation for medical domain

Now that we have the data, we can address our first research question: what happens we use news-trained segmenters on medical data?

## Task: Discourse segmentation for medical domain

- 1) What are difficulties of *news-trained* segmenters on *medical* data?
- 2) How do *features* of the segmenter impact the *type of errors* seen in medical?
- 3) What is the relationship between *annotator agreement* and *segmenter performance* for different types of medical data?

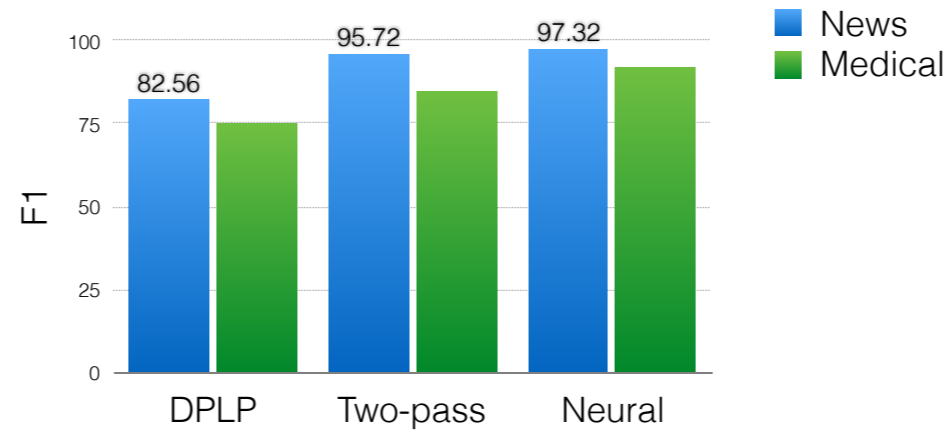
Now that we have the data, we can address our first research question: what happens we use news-trained segmenters on medical data?

## News-trained segmenters on medical?

- **DPLP**: uses features from syntactic and dependency parses for a linear support vector classifier (parser in Ji & Eisenstein, 2014)
- **Two-pass**: CRF segmenter that derives features from syntax parses but also uses global features to perform a second pass of segmentation (Feng & Hirst, 2014)
- **Neural**: neural BiLSTM-CRF model that uses ELMo embeddings (Wang et al., 2018)

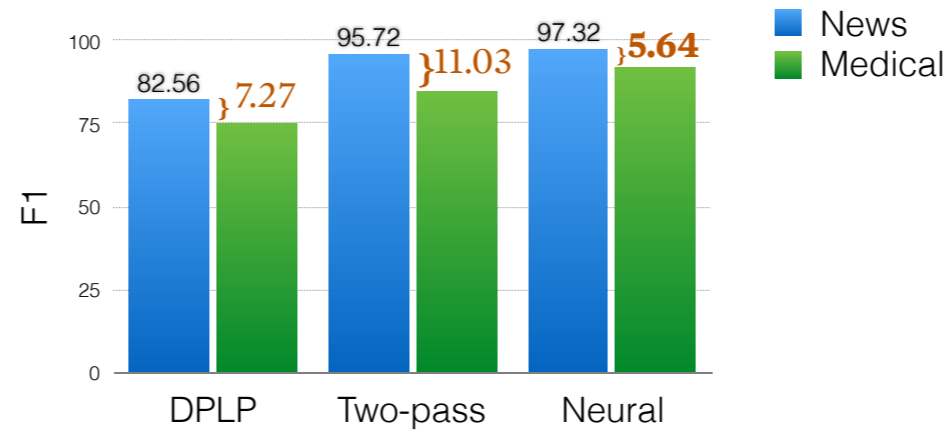
For automated segmenters, we choose the following three because they are publicly available and widely used. The first, DPLP uses features from syntactic and dependency parses for a linear SVC. Two-pass is a CRF segmenter also with features syntax trees but additionally uses global features to perform a second pass of segmentation. Finally, Neural is a neural BiLSTM-CRF that uses ELMo embeddings and this last achieves SOTA on segmented the RST-DT.

## News-trained segmenters on medical?



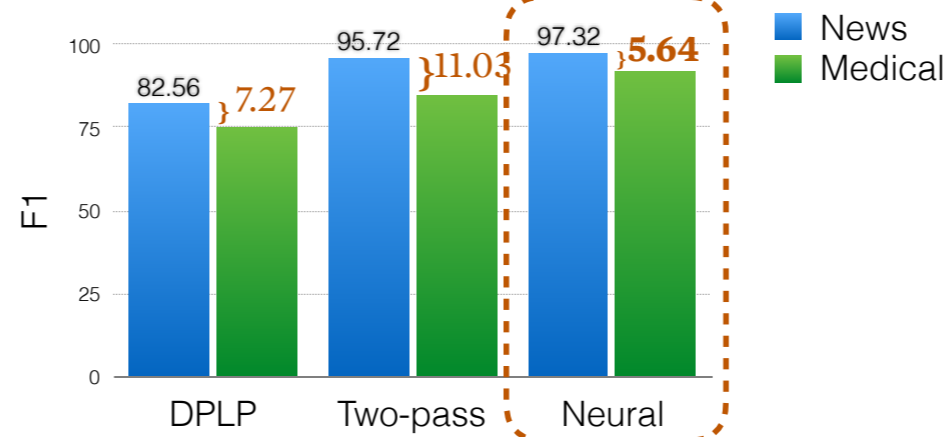
We segment the medical corpus and evaluate using F1 for matching all gold EDU boundaries. First for News in Blue we see that Neural does the best, which is expected since it achieves SOTA. However, we also see that Medical in green always does worse than News, no matter what segmenter you use. Because the gap is smallest with Neural parser, we focus on this parser for our error analysis.

## News-trained segmenters on medical?



We segment the medical corpus and evaluate using F1 for matching all gold EDU boundaries. First for News in Blue we see that Neural does the best, which is expected since it achieves SOTA. However, we also see that Medical in green always does worse than News, no matter what segmenter you use. Because the gap is smallest with Neural parser, we focus on this parser for our error analysis.

## News-trained segmenters on medical?



We segment the medical corpus and evaluate using F1 for matching all gold EDU boundaries. First for News in Blue we see that Neural does the best, which is expected since it achieves SOTA. However, we also see that Medical in green always does worse than News, no matter what segmenter you use. Because the gap is smallest with Neural parser, we focus on this parser for our error analysis.



# Error Type Categorization

We examine all the segmentation errors and attempt to categorize them into buckets or error types to understand the nature of the errors. First, we look at the most common error types in the News domain. Most of these are false positives where the segmenter inserted an EDU boundary that didn't belong there. The first is ambiguous lexical cue- usually the discourse connective "since" is a strong marker of an EDU boundary. But in this case there is no verbal element in "since the buy-out" so it's not a separate EDU.

# Error Type Categorization

Error Type

Example

---

ambiguous lexical cue [our performance]||since the buy-out makes it imperative]

We examine all the segmentation errors and attempt to categorize them into buckets or error types to understand the nature of the errors. First, we look at the most common error types in the News domain. Most of these are false positives where the segmenter inserted an EDU boundary that didn't belong there. The first is ambiguous lexical cue- usually the discourse connective "since" is a strong marker of an EDU boundary. But in this case there is no verbal element in "since the buy-out" so it's not a separate EDU.

# Error Type Categorization

Error Type

Example

---

ambiguous lexical cue [our performance ~~]~~ since the buy-out makes it imperative]

We examine all the segmentation errors and attempt to categorize them into buckets or error types to understand the nature of the errors. First, we look at the most common error types in the News domain. Most of these are false positives where the segmenter inserted an EDU boundary that didn't belong there. The first is ambiguous lexical cue- usually the discourse connective "since" is a strong marker of an EDU boundary. But in this case there is no verbal element in "since the buy-out" so it's not a separate EDU.

# Error Type Categorization

Error Type	Example
ambiguous lexical cue	[our performance <del>  </del> <u>since</u> the buy-out makes it imperative]
infinitival “to”	[the auto giants will move quickly <del>  </del> <u>to buy up</u> stakes]

The next error is with the infinitival “to” + verb. This has also been found to be problematic in prior work of Braud et al 2017. This construction can act either as a verbal complement or a clausal complement. In this example, “to buy up” is a complement of the verb “move” so it shouldn’t be segmented.

# Error Type Categorization

Error Type	Example
ambiguous lexical cue	[our performance <del>X</del> <u>since</u> the buy-out makes it imperative]
infinitival “to”	[the auto giants will move quickly <del>X</del> <u>to buy up</u> stakes]

The next error is with the infinitival “to” + verb. This has also been found to be problematic in prior work of Braud et al 2017. This construction can act either as a verbal complement or a clausal complement. In this example, “to buy up” is a complement of the verb “move” so it shouldn’t be segmented.

## Error Type Categorization

Error Type	Example
ambiguous lexical cue	[our performance <del>  </del> since the buy-out makes it imperative]
infinitival “to”	[the auto giants will move quickly <del>  </del> to buy up stakes]
annotator error	[you attempt to seize assets <del>  </del> related to the crime]


And lastly, annotator error is where the segmenter was right but the annotation was wrong. Here, “related to the crime” is a nominal postmodifier with a verbal element and should be treated as an embedded EDU.

## Error Type Categorization

Error Type	Example
ambiguous lexical cue	[our performance <del>  </del> <u>since</u> the buy-out makes it imperative]
infinitival “to”	[the auto giants will move quickly <del>  </del> <u>to buy up</u> stakes]
annotator error	[you attempt to seize assets <del>  </del> <u>related to the crime</u> ]

And lastly, annotator error is where the segmenter was right but the annotation was wrong. Here, “related to the crime” is a nominal postmodifier with a verbal element and should be treated as an embedded EDU.



# Error Type Categorization

Error Type	Example
ambiguous lexical cue	[our performance <del>X</del> since the buy-out makes it imperative]
infinitival “to”	[the auto giants will move quickly <del>X</del> to buy up stakes]
annotator error	[you attempt to seize assets  related to the crime]
punctuation	[the safety of placeboetine][ ( PB) hydrochloride capsules]

Moving to the medical domain, most errors are false negatives where the parser fails to detect an EDU boundary when there actually is one. The most frequent error types are first punctuation. If you recall from our “easy” segmentation exercise, parentheticals are always segmented. Here, the segmenter detects the open parenthesis but not the closing one. This suggests the segmenter is failing to learn this “hard” rule even though every parenthesis in RST-DT gets segmented.





# Error Type Categorization

Error Type	Example
ambiguous lexical cue	[our performance <del>X</del> since the buy-out makes it imperative]
infinitival “to”	[the auto giants will move quickly <del>X</del> to buy up stakes]
annotator error	[you attempt to seize assets  related to the crime]
punctuation	[the safety of placeboetine][ ( PB  hydrochloride capsules]







Moving to the medical domain, most errors are false negatives where the parser fails to detect an EDU boundary when there actually is one. The most frequent error types are first punctuation. If you recall from our “easy” segmentation exercise, parentheticals are always segmented. Here, the segmenter detects the open parenthesis but not the closing one. This suggests the segmenter is failing to learn this “hard” rule even though every parenthesis in RST-DT gets segmented.

# Error Type Categorization

Error Type	Example
ambiguous lexical cue	[our performance <del>X</del> since the buy-out makes it imperative]
infinitival “to”	[the auto giants will move quickly <del>X</del> to buy up stakes]
annotator error	[you attempt to seize assets  related to the crime]
punctuation	[the safety of placeboetine][ ( PB )  hydrochloride capsules]
end embedded EDU	[Studies][ confined to medical <b>professionals have</b> shown]





We talked before about embedded EDUs. In the medical domain, while segmenter detects the beginning of the embedded EDU, it often fails to detect the end. Interestingly, if you had a syntax tree you could clearly see where the end of that embedded EDU should be. Recall this parser uses ELMo embeddings which have been shown to learn syntax are too news-specific.

# Error Type Categorization

Error Type	Example
ambiguous lexical cue	[our performance  since the buy-out makes it imperative]
infinitival “to”	[the auto giants will move quickly  to buy up stakes]
annotator error	[you attempt to seize assets  related to the crime]
punctuation	[the safety of placeboetine][ ( PB  hydrochloride capsules]
end embedded EDU	[Studies][ confined to medical  professionals  have shown]


We talked before about embedded EDUs. In the medical domain, while segmenter detects the beginning of the embedded EDU, it often fails to detect the end. Interestingly, if you had a syntax tree you could clearly see where the end of that embedded EDU should be. Recall this parser uses ELMo embeddings which have been shown to learn syntax are too news-specific.

# Error Type Categorization

Error Type	Example
ambiguous lexical cue	[our performance <del>X</del> since the buy-out makes it imperative]
infinitival “to”	[the auto giants will move quickly <del>X</del> to buy up stakes]
annotator error	[you attempt to seize assets  related to the crime]
punctuation	[the safety of placeboetine][ ( PB  hydrochloride capsules]
end embedded EDU	[Studies][ confined to medical  professionals  have shown]
tokenization	[as identified in clinical trials.{8-11}It][is noteworthy]

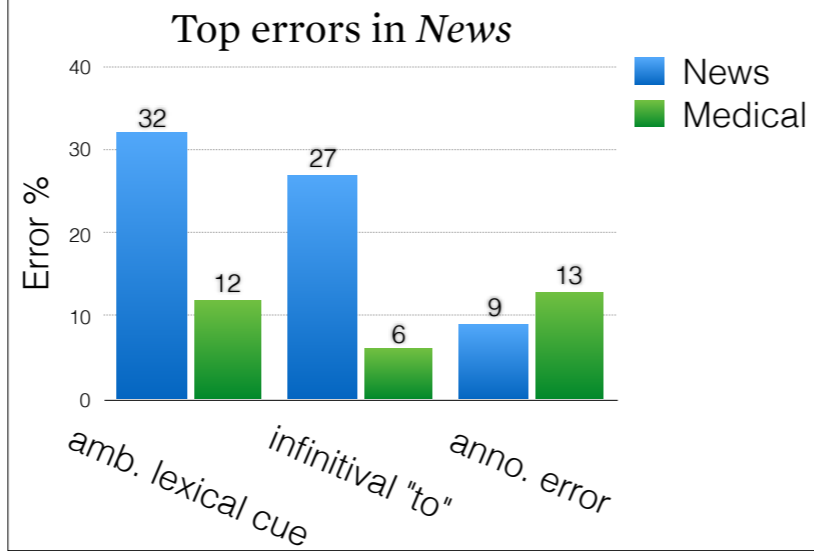
Finally, even before segmentation happens the tokenizer has issues correctly separating these tokens, which leads to these downstream errors. This segmenter uses spacy tokenizer which is also trained on news but is also really fast. Now that we’ve characterized what kind of errors happen in news and what kind in medical, let’s compare their distributions.

# Error Type Categorization

Error Type	Example
ambiguous lexical cue	[our performance <del>X</del> since the buy-out makes it imperative]
infinitival “to”	[the auto giants will move quickly <del>X</del> to buy up stakes]
annotator error	[you attempt to seize assets]  [related to the crime]
punctuation	[the safety of placeboetine][ ( PB ) ] hydrochloride capsules]
end embedded EDU	[Studies][ confined to medical professionals ] have shown]
tokenization	[as identified in clinical trials [ 8-11 ] It ] is noteworthy]

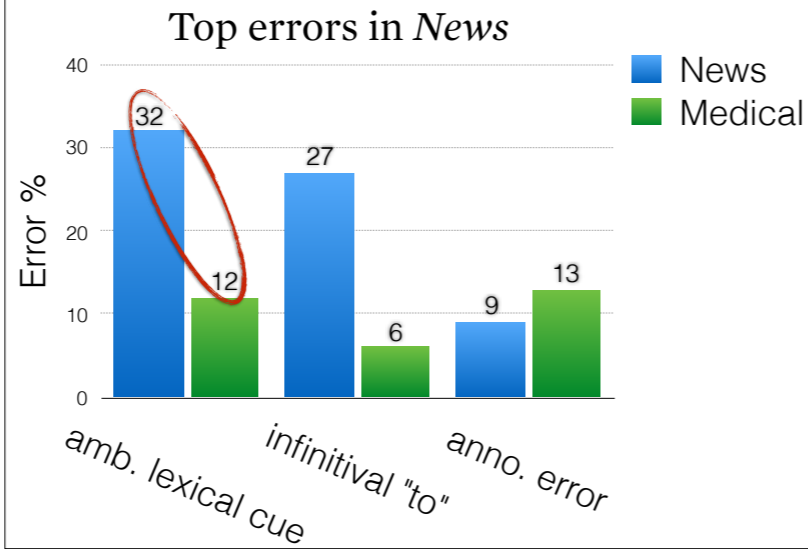
Finally, even before segmentation happens the tokenizer has issues correctly separating these tokens, which leads to these downstream errors. This segmenter uses spacy tokenizer which is also trained on news but is also really fast. Now that we’ve characterized what kind of errors happen in news and what kind in medical, let’s compare their distributions.

# Error Type Distribution



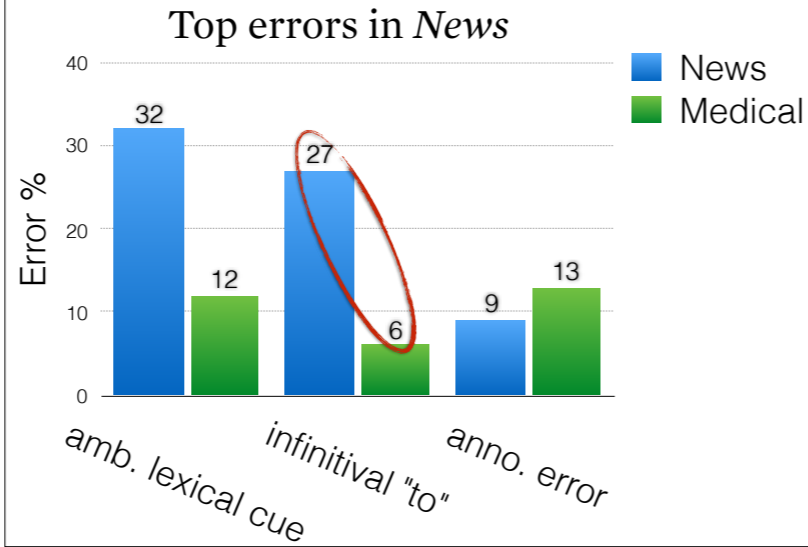
In news, the most common error type was the ambiguous lexical cue, which is also present in medical but not as frequent, which points to a difference in the way discourse connectives are used in the two domains.

# Error Type Distribution



In news, the most common error type was the ambiguous lexical cue, which is also present in medical but not as frequent, which points to a difference in the way discourse connectives are used in the two domains.

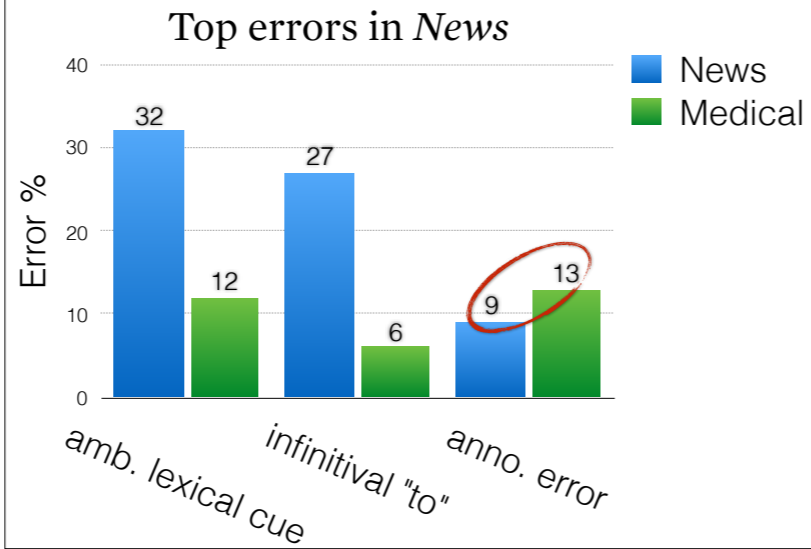
# Error Type Distribution



For infinitival to, it is considerably less frequent in medical in part because the to+verb construction is not as common.

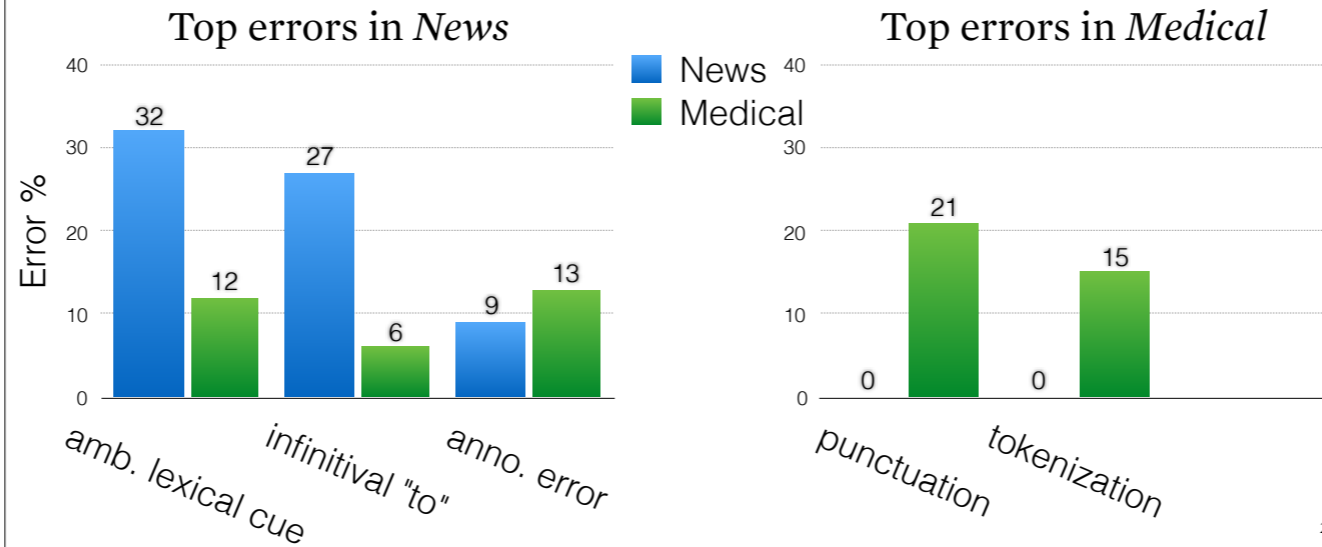


# Error Type Distribution



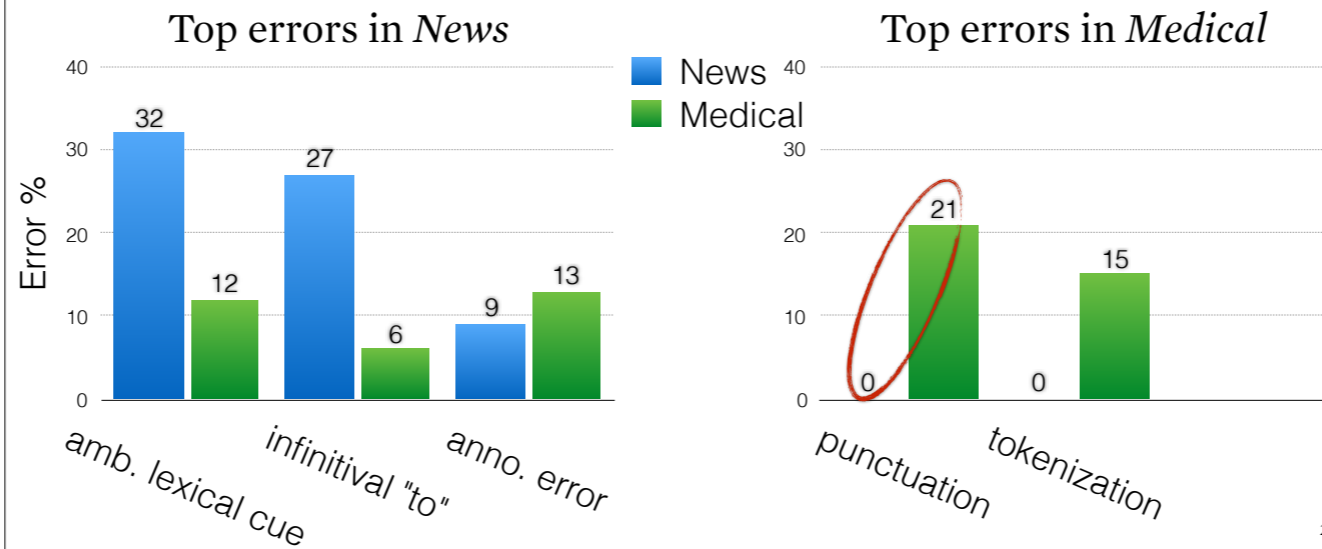
Finally the annotator error distribution shows that our error rate is very comparable to much larger-scale corpus that involved many expert linguists.

# Error Type Distribution



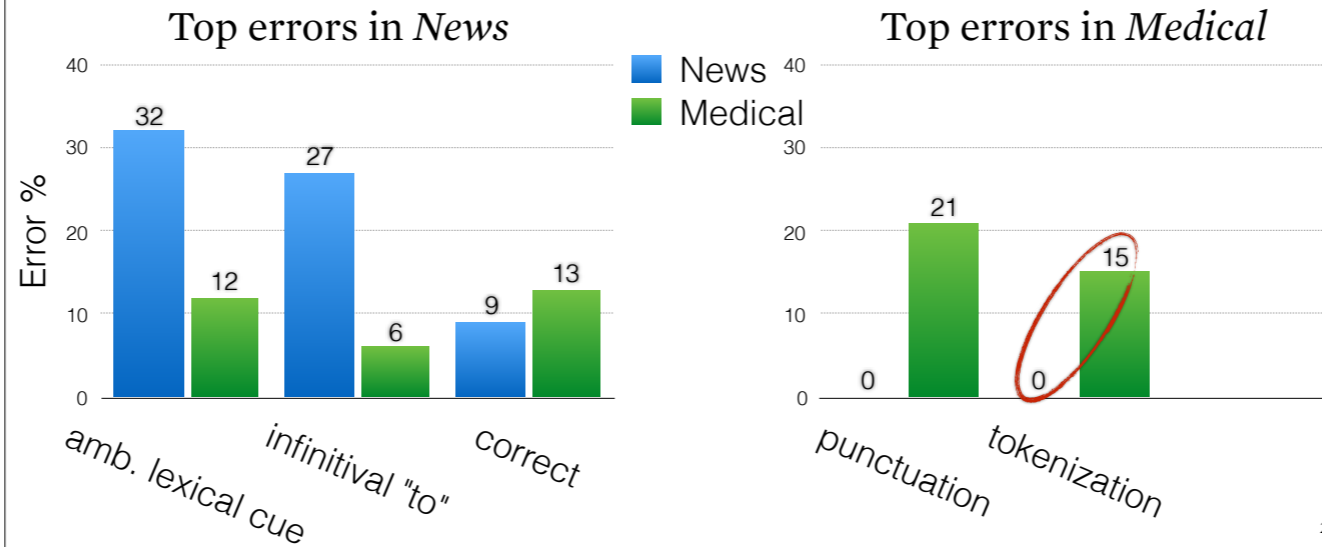
Next, we look at the distribution for the most common error types in medical. First, you'll notice that the top two error types in medical don't even occur in news. There are no issues with parentheses, and also no square brackets that the parser doesn't know what to do with.

# Error Type Distribution



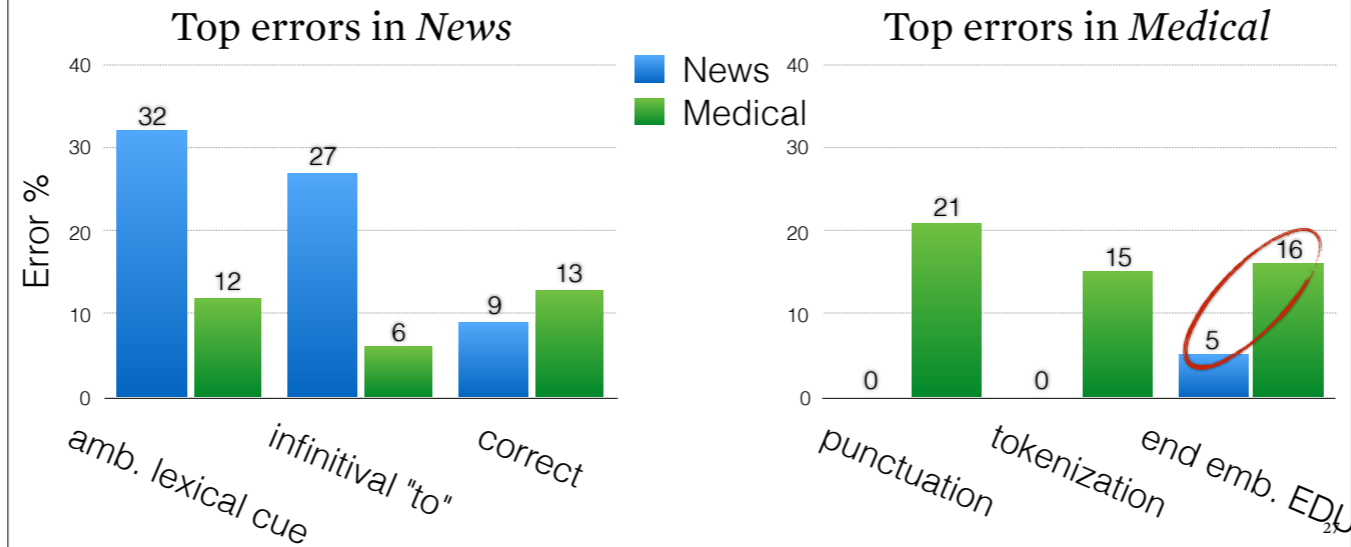
Next, we look at the distribution for the most common error types in medical. First, you'll notice that the top two error types in medical don't even occur in news. There are no issues with parentheses, and also no square brackets that the parser doesn't know what to do with.

# Error Type Distribution



There are no errors that can be traced back to tokenization issues.

# Error Type Distribution



And finally, hardly any problems detecting the end of embedded EDUs in news, but much harder in medical.

# News-trained segmenters on medical?

So we see a lot of the errors in the medical domain stem from formatting differences but also syntactic differences.

## News-trained segmenters on medical?

1) What are difficulties of *news-trained* segmenters on *medical* data?

***Formatting*** differences, ***syntactic*** differences

So we see a lot of the errors in the medical domain stem from formatting differences but also syntactic differences.

# Task: Discourse segmentation for medical domain

Our next question tries to understand whether features of the different segmenters could affect the types of errors we see in medical.



## Task: Discourse segmentation for medical domain

- 1) What are difficulties of *news-trained* segmenters on *medical* data?
- 2) How do *features* of the segmenter impact the *type of errors* seen in medical?
- 3) What is the relationship between *annotator agreement* and *segmenter performance* for different types of medical data?

Our next question tries to understand whether features of the different segmenters could affect the types of errors we see in medical.

## Features of segmenter?

end embedded EDU [Studies][ confined to medical professionals][have shown]  
tokenization [as identified in clinical trials][8-11][It][is noteworthy]

- Many medical errors could be resolved by syntax or better tokenizer

Remember when we talked about finding the end of an embedded EDU. If we only had a parse tree, it would be obvious. Or the tokenization errors. What if we ditched the spacy parser? What if we had a segmenter that used a good tokenizer like Stanford Core NLP Tokenizer, and features from Stanford Core NLP syntactic trees? I'm describing the Two-pass segmenter.

## Features of segmenter?

end embedded EDU [Studies][ confined to medical professionals][have shown]  
tokenization [as identified in clinical trials][8-11][It][is noteworthy]

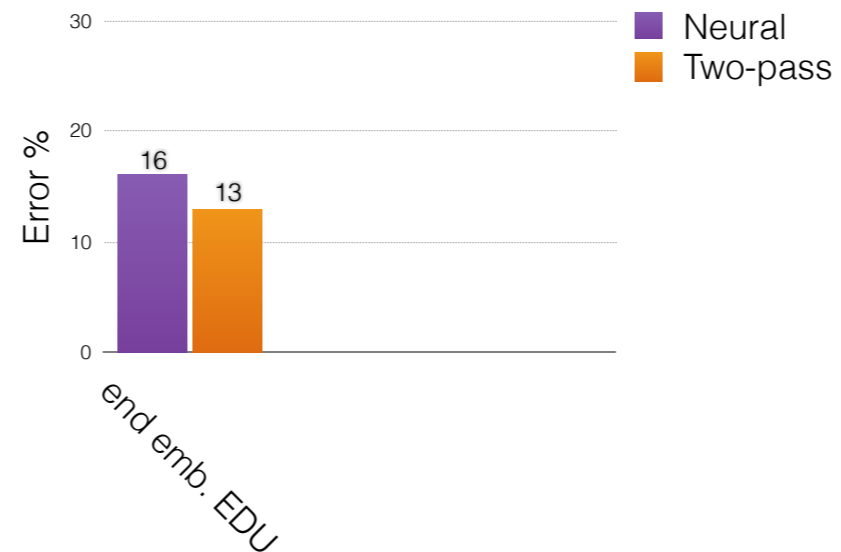
- Many medical errors could be resolved by syntax or better tokenizer
- Could a segmenter with ***syntactic parse trees*** or a ***better tokenizer*** help?

Remember when we talked about finding the end of an embedded EDU. If we only had a parse tree, it would be obvious. Or the tokenization errors. What if we ditched the spacy parser? What if we had a segmenter that used a good tokenizer like Stanford Core NLP Tokenizer, and features from Stanford Core NLP syntactic trees? I'm describing the Two-pass segmenter.

# Neural vs. Two-pass

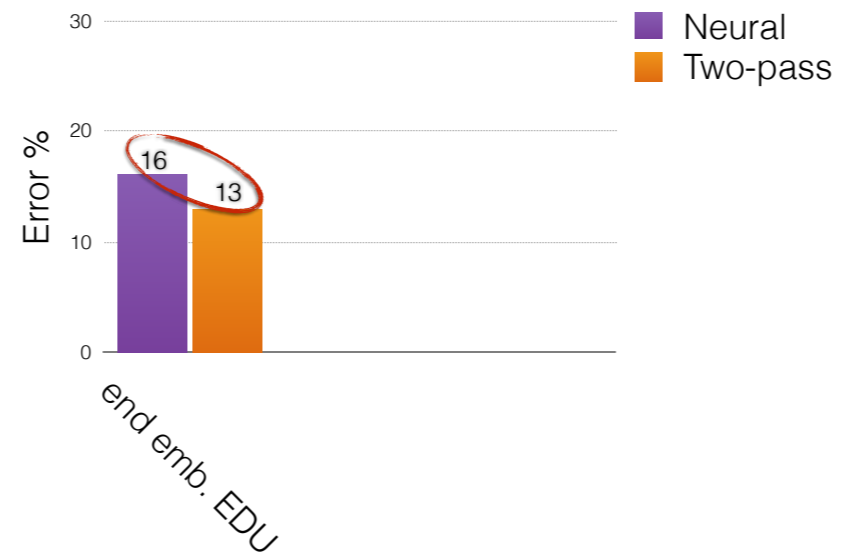
With access to the syntax trees, the error distribution doesn't change much. But these aren't gold trees, so maybe even Stanford Core NLP has trouble parsing these out-of-domain sentences. We manually inspected these errors and found only about half were incorrectly parsed, so we suspect even gold trees might not be that helpful, which is in line with Braud et al 2017.

## Neural vs. Two-pass



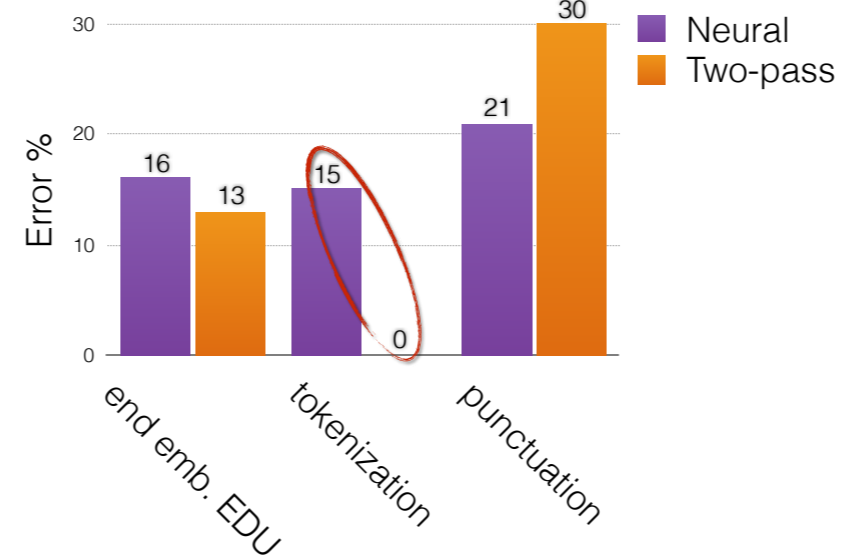
With access to the syntax trees, the error distribution doesn't change much. But these aren't gold trees, so maybe even Stanford Core NLP has trouble parsing these out-of-domain sentences. We manually inspected these errors and found only about half were incorrectly parsed, so we suspect even gold trees might not be that helpful, which is in line with Braud et al 2017.

## Neural vs. Two-pass



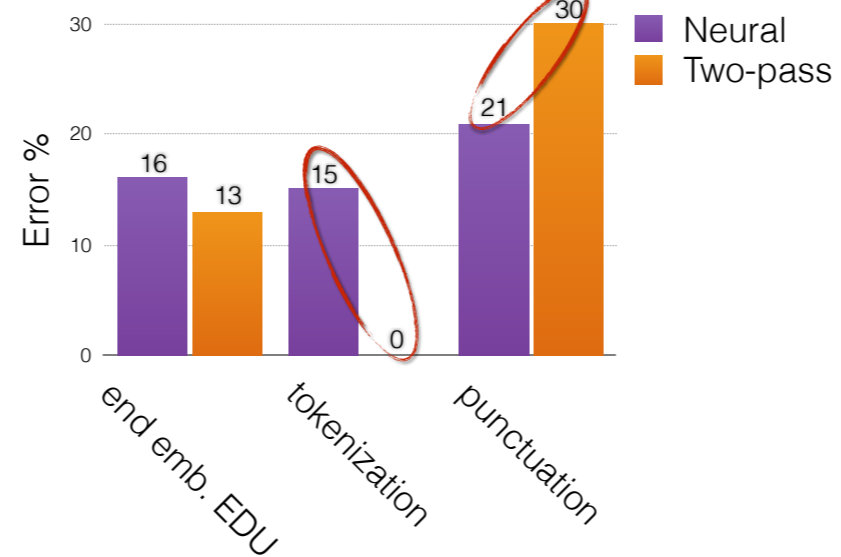
With access to the syntax trees, the error distribution doesn't change much. But these aren't gold trees, so maybe even Stanford Core NLP has trouble parsing these out-of-domain sentences. We manually inspected these errors and found only about half were incorrectly parsed, so we suspect even gold trees might not be that helpful, which is in line with Braud et al 2017.

## Neural vs. Two-pass



For tokenization, the Two-pass segmenter makes 0 of these kinds of errors. But we also see that it makes more punctuation errors. So could it be that we're just shifting errors from one type to another? We're able to identify those tokens with punctuations so now we mess up on them more often. We tried using the neural parser with pre-segmented text (that is, bypassing the spacy tokenizer) and we get only a slight bump in performance (F1 +0.36) so the answer is.

## Neural vs. Two-pass



For tokenization, the Two-pass segmenter makes 0 of these kinds of errors. But we also see that it makes more punctuation errors. So could it be that we're just shifting errors from one type to another? We're able to identify those tokens with punctuations so now we mess up on them more often. We tried using the neural parser with pre-segmented text (that is, bypassing the spacy tokenizer) and we get only a slight bump in performance (F1 +0.36) so the answer is.



# Features of segmenter?

In conclusion, by comparing these two segmenters we saw that syntax trees did not help and a better tokenizer helps, but only a little bit.

## Features of segmenter?

2) How do **features** of the segmenter impact the **type of errors** seen in medical?

Syntax trees **do not help**, and a better tokenizer only helps **a little**.

In conclusion, by comparing these two segmenters we saw that syntax trees did not help and a better tokenizer helps, but only a little bit.

# Task: Discourse segmentation for medical domain

Finally, we compare the inter-annotator agreement and the segmenter performance by what section of medical data we are segmenting.

## Task: Discourse segmentation for medical domain

- 1) What are difficulties of *news-trained* segmenters on *medical* data?
- 2) How do *features* of the segmenter impact the *type of errors* seen in medical?
- 3) What is the relationship between *annotator agreement* and *segmenter performance* for different types of medical data?

Finally, we compare the inter-annotator agreement and the segmenter performance by what section of medical data we are segmenting.

## Annotator agreement vs. Performance?

Section	Kappa	F1	Tokens
Summary	1.00	100	35
Introduction	0.96	86.58	258
Results	0.93	91.74	354
Abstract	0.89	95.08	266
Methods	0.86	92.99	417
Discussion	0.84	89.03	365

Here I list the section, the inter annotator agreement as measured by the kappa coefficient, the performance of the automated segmenter in F1, and the average number of tokens in that section. For very short sections like Summary, both humans and automated segmenter do very well.

## Annotator agreement vs. Performance?

Section	Kappa	F1	Tokens
Summary	1.00	100	35
Introduction	0.96	86.58	258
Results	0.93	91.74	354
Abstract	0.89	95.08	266
Methods	0.86	92.99	417
Discussion	0.84	89.03	365

Here I list the section, the inter annotator agreement as measured by the kappa coefficient, the performance of the automated segmenter in F1, and the average number of tokens in that section. For very short sections like Summary, both humans and automated segmenter do very well.

## Annotator agreement vs. Performance?

Section	Kappa	F1	Tokens
Summary	1.00	100	35
Introduction	0.96	86.58	258
Results	0.93	91.74	354
Abstract	0.89	95.08	266
Methods	0.86	92.99	417
Discussion	0.84	89.03	365

The Introduction, which usually has simple discourse like the example we saw at the beginning, is also easy for humans. But the Introduction also has lots of citations which makes it more difficult for the automated segmenter.

## Annotator agreement vs. Performance?

Section	Kappa	F1	Tokens
Summary	1.00	100	35
Introduction	0.96	86.58	258
Results	0.93	91.74	354
Abstract	0.89	95.08	266
Methods	0.86	92.99	417
Discussion	0.84	89.03	365

Finally, the discussion has more complex discourse. Recall the example of the single sentence with 6 edus. This sections is hard for both humans and segmenters.



# Annotator agreement vs. Performance?

To summarize, there are two factors that affect inter annotator agreement and segmenter performance: text length, where very short texts are easy for both. And discourse complexity, where more complex is harder for both humans and segmenters.

## Annotator agreement vs. Performance?

3) What is the relationship between **annotator agreement** and **segmenter performance** for different types of medical data?

- **Text length:** Very short texts are easy for both humans and segmenter
- **Discourse complexity:** More complex discourse is harder for both humans and segmenter

To summarize, there are two factors that affect inter annotator agreement and segmenter performance: text length, where very short texts are easy for both. And discourse complexity, where more complex is harder for both humans and segmenters.

## Next steps:

Because we've presented a small corpus, the natural next question is how to expand it. First, you should use the Neural segmenter as a first pass. Next, focus manual effort on the more discourse-complex sections like Discussion. During training, use the medical specific tokenizer and word embeddings. But an important question to ask which we'd love to address once we have trees, is what effect do these errors have on the downstream tasks of span, nuclearity and relations labeling? And beyond that, what effect do they have on downstream tasks that people in the medical domain care about, like labeling sentences that describe results versus other information?

## Next steps:

- Use **Neural** segmenter for first pass
- Focus **manual** annotation on more **discourse-complex** sections (e.g., *Discussion*)
- During training: use **medical-specific** preprocessing + word embeddings

Because we've presented a small corpus, the natural next question is how to expand it. First, you should use the Neural segmenter as a first pass. Next, focus manual effort on the more discourse-complex sections like Discussion. During training, use the medical specific tokenizer and word embeddings. But an important question to ask which we'd love to address once we have trees, is what effect do these errors have on the downstream tasks of span, nuclearity and relations labeling? And beyond that, what effect do they have on downstream tasks that people in the medical domain care about, like labeling sentences that describe results versus other information?

# Thank you!

**Corpus & code:**

<https://github.com/elisaF/news-med-segmentation>

`elisa@ferracane.com`

Thank you and here is the url for our corpus and code for analyses.