

# Grounded Complex Task Segmentation for Conversational Assistants

Rafal Ferreira, David Semedo, João Magalhães

NOVA University of Lisbon

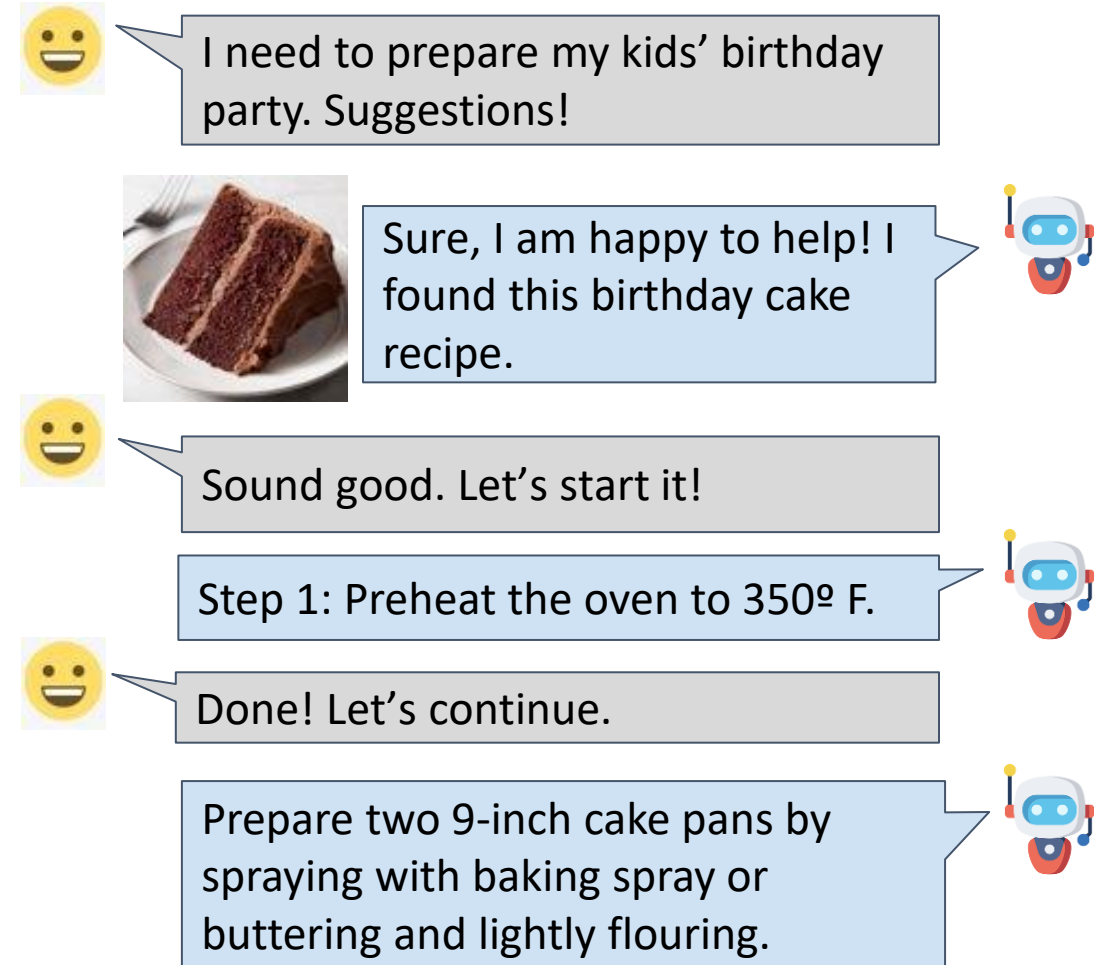
NOVA LINCS

Lisbon, Portugal



# Conversational Task Assistants

- Conversational Task Assistants (**CTAs**) guide users through everyday tasks such as cooking and DIY.
- Their main functions include:
  - Understanding user intentions.
  - Communicate instructions in a **structured** and **well-paced** manner.
  - Collaborate closely with users, facilitating the execution of the task.



I need to prepare my kids' birthday party. Suggestions!

Sure, I am happy to help! I found this birthday cake recipe.

Sound good. Let's start it!

Step 1: Preheat the oven to 350° F.

Done! Let's continue.

Prepare two 9-inch cake pans by spraying with baking spray or buttering and lightly flouring.

...

# Motivation and Challenges

- Online instructional texts are often **suboptimal** for conversational assistants due to differences between screen and voice-based interactions.
- Recipes show the need for decomposing text into dialogue-suited steps:

**Step N** - Discard excess oil from pan and then add green onion, ginger, garlic, chilis, star anise, cinnamon, and bay leaves. Next, add soy sauce, wine, tofu, and water. Reduce heat to low and cook for approx. 1 – 2 hrs. until volume has reduced by half and sauce has thickened.

Long, complex, hard to follow

**Step N** - Discard excess oil from pan and then add green onion, ginger, garlic, chilis, star anise, cinnamon, and bay leaves.

**Step N+1** - Next, add soy sauce, wine, tofu, and water. Reduce heat to low and cook for approx. 1 – 2 hrs. until volume has reduced by half and sauce has thickened.

- **Segment** each step into **manageable** pieces of information, balancing **complexity**.

# Contributions

- Creation of the **ConvRecipes** corpus:
  - Demonstrated differences between web-based and dialogue-suited instructions.
- Proposed and evaluated several methods to capture **conversational-instructional** segmentation patterns.
- Results showed that the best model **improved the conversational structure** of 86% of the evaluated tasks.



# Structuring Conversational Tasks

# Structuring Conversational Tasks

- **Hypothesis:**
  - Online recipe instructions are **not suitable** for conversational assistants.
- **Methodology:**
  - Convert instructions from a **reading structure** into a **conversational structure**.
- **Steps:**
  1. **Collected** task instructions (e.g., recipes).
  2. Manually **curate** the instructions into a conversational setting.
  3. Asked users to annotate the relevance of conversational instruction **traits**.
  4. Analyzed **linguistic characteristics** of reading vs. conversational task instructions.
  5. **Modeled** task segmentation using several methods.

# A Conversational-Tasks Corpus

- Lack of explicit corpora for studying **conversational task segmentation**:
  - Section-based segmentation – e.g., WikiSection (Arnold et al., 2019) – non-instructional text.
  - Generative/Re-writing - e.g., Task2Dial (Strathearn et al., 2021) – prone to model changing the task (hallucination).
- We introduce the **ConvRecipes** corpus for **grounded** recipe segmentation for a **conversational** setting.

# Collection Task Instructions

- Collected recipes from popular recipes websites.
- Recipes are self-contained texts with multiple ordered steps in English.
- Filtered out:
  - Recipes with fewer than three steps.
  - Near-duplicates removal.



## Bold & Smoky Tomato Chipotle Soup

SERVINGS: 6

PREPPING TIME: 15 MIN

COOKING TIME: 30 MIN

### Ingredients

2 tablespoons unsalted butter  
 2 small onions, finely chopped  
 2 garlic cloves, minced  
 1 1/2 tablespoons tomato paste  
 3 tbsp sugar  
 1/4 cup TABASCO® Chipotle Sauce (adjust depending on your spice preference)  
 800g canned roasted tomatoes  
 2/3 cup cooking sherry  
 1/3 cup chicken stock  
 1/2 teaspoon dried basil  
 1/4 teaspoon salt  
 1/4 teaspoon pepper  
 3/4 cup heavy cream  
 freshly snipped basil for garnish

### Directions

1. Heat olive oil over medium-low heat. Add onion and cook, stirring occasionally, until softened and translucent – about 3 minutes.
2. Add the garlic, sugar, tomato paste, and TABASCO® Chipotle Sauce. Cook for 2 more minutes.
3. Add the fire-roasted tomatoes, sherry, chicken stock, dried basil, salt, pepper and stir. Bring to a boil, then reduce to a simmer and cover, cooking for 10 minutes.
4. Turn off the heat and use an immersion blender to purée the soup (or if using a regular blender, let the soup cool for at least 10 minutes first to avoid an explosion). Using a high-speed blender will give you a smoother soup.
5. Once the soup is blended, add the cream and return the soup to a slight simmer, just to heat it through. Enjoy warm! Serve with fresh bread.



# Annotating Dialogue-suited Steps

- **User study** with 8 annotators:
  - 6 male, 2 female, all with Computer Science MSc. or Ph.D.
  - Annotators experienced with conversational assistants and cooking apps.
- **Task:**
  - Modify original recipes to make them **dialog-suited** by adding or removing segmentations.
  - Segmentation **grounded on the original recipe**, avoiding rewriting mistakes.

## **Title:** Baked Bananas Recipe

---

### Web-based Recipe

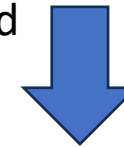
---

**Step 1:** Preheat oven to 190 degrees C. Spray a baking dish with cooking spray.

**Step 2:** Arrange banana halves in the prepared baking dish. Drizzle maple syrup over bananas and top with cinnamon. Bake in the oven until heated through, 10-15 minutes.

---

From Web-Based



To Dialogue-suited

---

### Dialogue-suited Recipe

---

**Step 1:** Preheat oven to 190 degrees C.

**Step 2:** Spray a baking dish with cooking spray.

**Step 3:** Arrange banana halves in the prepared baking dish. Drizzle maple syrup over bananas and top with cinnamon.

**Step 4:** Bake in the oven until heated through 10-15 minutes.

---

# On the Traits of a Conversational Step (1)

- Gain insights into **user preferences** for this task.
- Annotators were tasked with quantifying the importance of traits in a **1 to 5 scale**.
- Considered traits:
  - **Complexity** - was the complexity of the step an important factor?
  - **Clarity** - was the information clear, organized, and well delimited in each step?
  - **Step Length & # Steps** – did the length and the total number of steps matter?
  - **Ability to Parallelize** - should it allow the user to parallelize steps?
  - **Naturalness** - was the naturalness of each step important?

# On the Traits of a Conversational Step (2)

- All traits hold some degree of importance.
- Most important traits are **complexity** and **length** of the steps.
  - Highlights the critical nature of maintaining a balance in the information provided to the user.
- Naturalness and the ability to perform parallel tasks were deemed less important.
- Users might prioritize **concise, straightforward** steps over language naturalness and multitasking capabilities.

Conversational-Step Trait	Importance
(1) Complexity	4.5
(2) Step Length & #Steps	4.2
(3) Clarity	3.8
(4) Naturalness	3.6
(5) Ability to Parallelize Tasks	3.4

Table 1: Trait importance on a 1 to 5 scale. A higher value represents higher importance.

# Reading-suited vs Dialog-suited

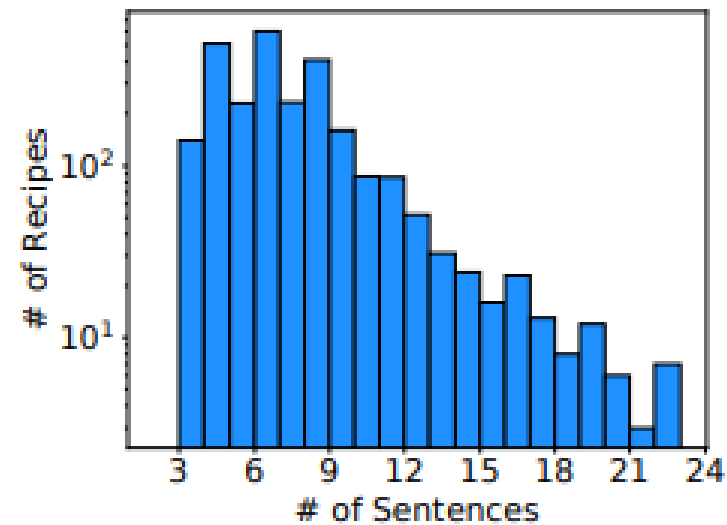
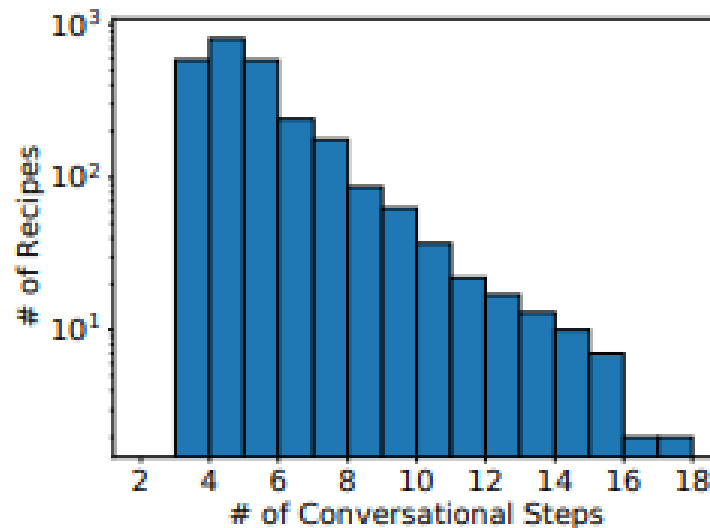
- A total of 300 recipes were annotated
  - 59 (20.7%) recipes remained unchanged.
  - 241 (**80.3%**) had at least one new step added.
- **Contrast** between original and dialog instructions:
  - Reading instructions often **lack critical segmentations**.
  - Preference for **shorter segments** with fewer actions.
  - Findings align with the traits analysis.

	Reading	Dialog
Avg. # Tokens	135	
Avg. # Sentences	9.3	
Avg. # Steps	3.80	5.85
Avg. # Tokens step	35.44	23.03
Avg. # Sents. step	2.44	1.59
Avg. # Verbs step	4.23	2.75
Avg. # Nouns step	9.92	6.44

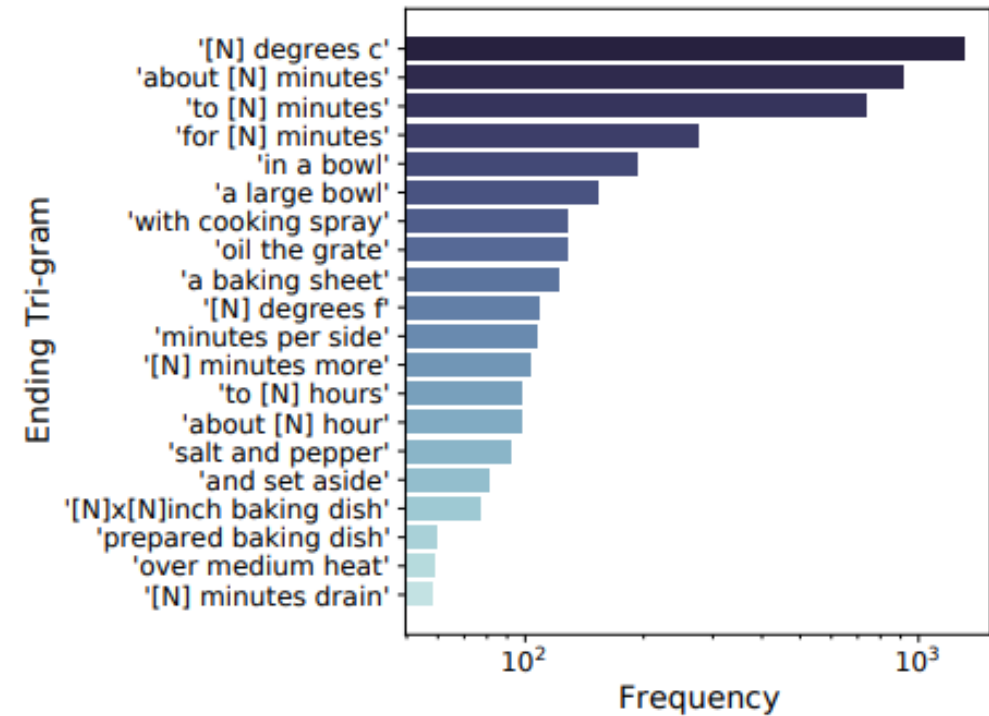
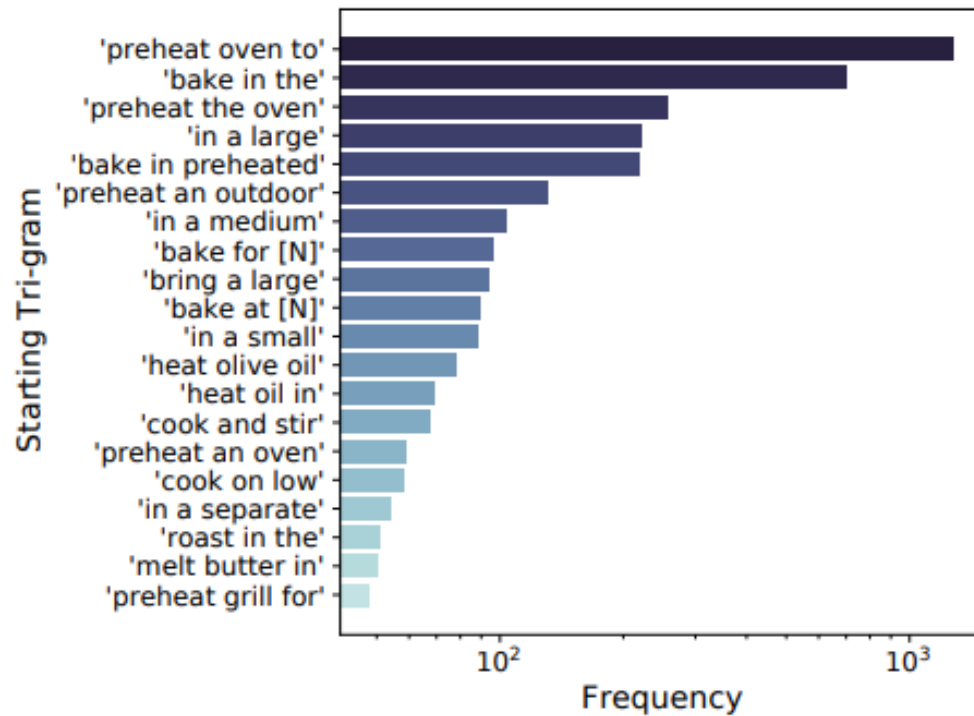
Table 2: Comparison between the 300 original reading-based recipes and the manually annotated set.

# Linguistic Style of Conversational-Steps (1)

- In the recipes domain there exists a substantial **variability**:



# Linguistic Style of Conversational-Steps (2)



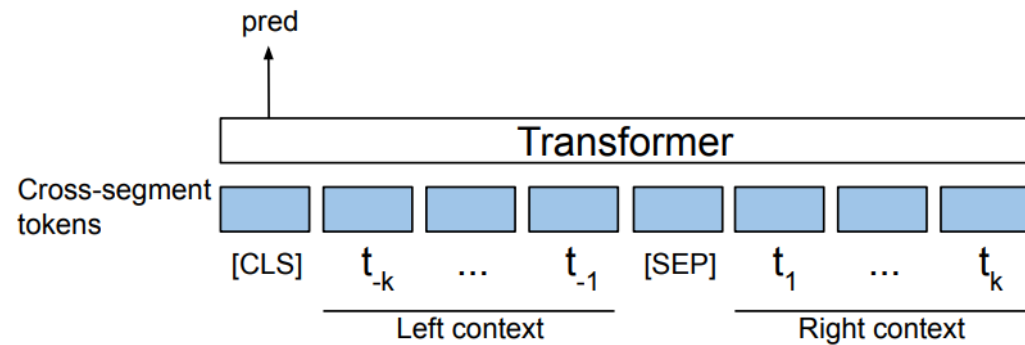
- Top-20 **starting** and **ending tri-grams** reveals insights into segmentation behavior:
  - **Temperature** mentions (e.g., "preheat oven to")
  - **Time-aware** mentions (e.g., "for [N] minutes")
  - Shows the **sequential aspect** of recipes and instructional text.

# Corpus Processing

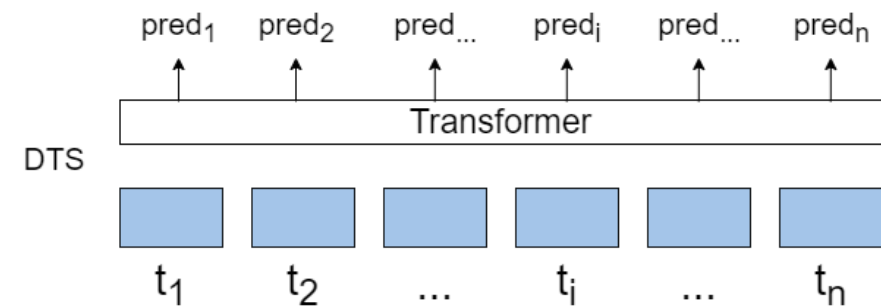
- Annotating recipes is labor-intensive and costly.
  - Test set comprises 300 manually annotated recipes.
  - Training and validation automatically created using statistics of the annotated set.
  - Non-annotated recipes utilize original step information as ground-truth labels.
- Dataset includes:
  - ~2000 for training, ~400 for validation, and 300 for testing.
- **The task:**
  - Given all steps concatenated into an unstructured text.
  - Develop methods to identify and segment tasks into **dialogue-suited steps**.

# Dialog-Task Structuring Transformer (DTS)

- The model is fed the **complete recipe**, allowing it to create contextualized token representations of the entire text.
  - Is able to consider the **position** of each token relative to all other tokens.
- DTS contrasts with sentence-based embedding models that predict segments per sentence.



Cross-Segment (Lukasik et al., 2020)



Dialog-Task Structuring Transformer (Ours)



# Dialog-Task Structuring Transformer (DTS)

- The **Transformer** model serves as the backbone.
- A linear layer followed by a *softmax*, is applied to the embedding of each **segment identifying token** ( $emb_t$ ).
- Returns the probability of a token ( $t$ ) being a *segmentation token*.

$$P_{seg}(t_i) = softmax(FFNN(emb_{[t_i]})), \quad (1)$$

- Model is trained with the **cross-entropy** loss, calculated between the model's predictions ( $\hat{y}$ ) and the segmentation labels ( $y$ ).

$$L_{CE} = y \cdot \log \hat{y} + (1 - y) \cdot \log (1 - \hat{y}), \quad (2)$$



# Experiments and Results

# Metrics

- **Exact match** metrics:
  - Segment predicted by the model is the same as in the ground-truth.
  - **Precision, Recall, and F-score.**
- Text segmentation metric  $P_k$  (Beeferman et al., 1999):
  - Slides a window of size  $K$ , which returns 0 if the sentences are in the same segment as the ground truth and 1 otherwise.
  - A lower value of  $P_k$  indicates a better model.

# Baselines and DTS Models

- Unsupervised methods using Spacy for sentence identification:
  - **Rand<sub>p</sub>** - where  $p$  is the probability of segmenting a sentence.
  - **Every<sub>n</sub>** – segment after  $n$  consecutive sentences.
- **TextTiling** - An early text segmentation method based on lexical co-occurrence.
- **Cross-Segment** (Lukasik et al., 2020) - Utilizes a BERT-Base model with a classification head to predict whether a pair of input sentences should be segmented.
- **DTS (ours)** - Backbone built upon:
  - BERT (Devlin et al., 2018) - encoder-only model
  - T5 (Raffel et al., 2019) – in both enc-dec and enc-only setting

# General Results

- Unsupervised baselines have **low precision** ( $\leq 62\%$ ).
- $\text{Every}_1$  achieves decent  $P_k$  and high recall due to breaking at every sentence.
- TextTiling performs poorly since it relies on lexical overlap.
- CrossSeg is an improvement over the unsupervised baselines.
- **DTS consistently outperforms** all baselines highlighting the importance of **token-level step relations**.

	Model	# Params	$P_k \downarrow$	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$
Baselines	Rand <sub>0.5</sub>	-	$35.4 \pm 0.3$	$59.9 \pm 0.5$	$49.7 \pm 0.8$	$51.7 \pm 0.6$
	Rand <sub>0.75</sub>	-	$28.3 \pm 0.5$	$61.2 \pm 0.4$	$75.0 \pm 0.9$	$65.2 \pm 0.6$
	Every <sub>1</sub>	-	23.3	60.9	<b>98.8</b>	73.8
	Every <sub>2</sub>	-	37.9	59.6	37.9	44.9
	TextTiling	-	28.4	58.7	67.7	61.4
	CrossSeg	110 M	$19.5 \pm 0.4$	$77.5 \pm 0.9$	$79.5 \pm 1.6$	$76.5 \pm 0.4$
Dialogue Task Segmenter (DTS)	BERT-Base (All*)	110 M	$22.5 \pm 0.3$	<b><math>93.4 \pm 0.1</math></b>	$58.7 \pm 0.4$	$69.6 \pm 0.4$
	BERT-Base	110 M	$19.1 \pm 0.4$	$75.8 \pm 0.7$	$83.6 \pm 0.7$	$77.5 \pm 0.4$
	BERT-Large	340 M	$18.4 \pm 0.2$	$77.0 \pm 1.7$	$83.6 \pm 2.8$	$78.1 \pm 0.5$
	T5-Base (Enc-only)	110 M	$17.7 \pm 0.2$	$77.9 \pm 0.7$	$84.2 \pm 0.5$	$79.0 \pm 0.1$
	T5-Base (Enc-Dec)	220 M	$18.1 \pm 0.6$	$77.9 \pm 0.3$	$82.9 \pm 1.6$	$78.5 \pm 0.8$
	T5-Large (Enc-only)	335 M	$18.1 \pm 0.2$	$77.4 \pm 0.4$	$84.1 \pm 0.4$	$78.6 \pm 0.3$
	T5-Large (Enc-Dec)	770 M	$17.7 \pm 0.2$	$79.1 \pm 0.8$	$81.9 \pm 0.9$	$78.5 \pm 0.2$
	T5-3B (Enc-only)	1.5 B	<b><math>17.0 \pm 0.4</math></b>	$78.3 \pm 1.0$	$85.9 \pm 0.9$	<b><math>80.0 \pm 0.2</math></b>

Table 3: Results on the ConvRecipes’s test set from an average of 3 runs per model. *All\** indicates that the model was trained on the set of all recipes crawled, in their original form.

# Importance of Conversational-Aware Corpora

- Two BERT-Base models trained with:
  - All crawled raw recipes (All\*)
  - ConvRecipes training set
- BERT-Base (All\*):
  - Highest precision and lowest recall.
  - Due to fewer breaks in its training samples.
- Results highlight importance of **training with conversational data** with a 15%  $P_k$  improvement.

	Model	# Params	$P_k \downarrow$	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$
Baselines	Rand <sub>0.5</sub>	-	35.4 $\pm$ 0.3	59.9 $\pm$ 0.5	49.7 $\pm$ 0.8	51.7 $\pm$ 0.6
	Rand <sub>0.75</sub>	-	28.3 $\pm$ 0.5	61.2 $\pm$ 0.4	75.0 $\pm$ 0.9	65.2 $\pm$ 0.6
	Every <sub>1</sub>	-	23.3	60.9	<b>98.8</b>	73.8
	Every <sub>2</sub>	-	37.9	59.6	37.9	44.9
	TextTiling	-	28.4	58.7	67.7	61.4
	CrossSeg	110 M	19.5 $\pm$ 0.4	77.5 $\pm$ 0.9	79.5 $\pm$ 1.6	76.5 $\pm$ 0.4
Dialogue Task Segmenter (DTS)	BERT-Base (All*)	110 M	22.5 $\pm$ 0.3	<b>93.4 <math>\pm</math> 0.1</b>	58.7 $\pm$ 0.4	69.6 $\pm$ 0.4
	BERT-Base	110 M	19.1 $\pm$ 0.4	75.8 $\pm$ 0.7	83.6 $\pm$ 0.7	77.5 $\pm$ 0.4
	BERT-Large	340 M	18.4 $\pm$ 0.2	77.0 $\pm$ 1.7	83.6 $\pm$ 2.8	78.1 $\pm$ 0.5
	T5-Base (Enc-only)	110 M	<u>17.7 <math>\pm</math> 0.2</u>	77.9 $\pm$ 0.7	84.2 $\pm$ 0.5	<u>79.0 <math>\pm</math> 0.1</u>
	T5-Base (Enc-Dec)	220 M	18.1 $\pm$ 0.6	77.9 $\pm$ 0.3	82.9 $\pm$ 1.6	78.5 $\pm$ 0.8
	T5-Large (Enc-only)	335 M	18.1 $\pm$ 0.2	77.4 $\pm$ 0.4	84.1 $\pm$ 0.4	78.6 $\pm$ 0.3
	T5-Large (Enc-Dec)	770 M	<u>17.7 <math>\pm</math> 0.2</u>	<u>79.1 <math>\pm</math> 0.8</u>	81.9 $\pm$ 0.9	78.5 $\pm$ 0.2
	T5-3B (Enc-only)	1.5 B	<b>17.0 <math>\pm</math> 0.4</b>	<u>78.3 <math>\pm</math> 1.0</u>	<u>85.9 <math>\pm</math> 0.9</u>	<b>80.0 <math>\pm</math> 0.2</b>

Table 3: Results on the ConvRecipes’s test set from an average of 3 runs per model. All\* indicates that the model was trained on the set of all recipes crawled, in their original form.

# Encoder vs Encoder-Decoder Backbones

- With similar parameters T5 **outperforms** BERT.
- T5:
  - T5-Base Enc-Dec small decrease.
  - T5-Large Enc-Dec small increase.
- Results suggests that using the decoder part may not be necessary for this task.

	Model	# Params	$P_k \downarrow$	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$
Baselines	Rand <sub>0.5</sub>	-	35.4 $\pm$ 0.3	59.9 $\pm$ 0.5	49.7 $\pm$ 0.8	51.7 $\pm$ 0.6
	Rand <sub>0.75</sub>	-	28.3 $\pm$ 0.5	61.2 $\pm$ 0.4	75.0 $\pm$ 0.9	65.2 $\pm$ 0.6
	Every <sub>1</sub>	-	23.3	60.9	<b>98.8</b>	73.8
	Every <sub>2</sub>	-	37.9	59.6	37.9	44.9
	TextTiling	-	28.4	58.7	67.7	61.4
	CrossSeg	110 M	19.5 $\pm$ 0.4	77.5 $\pm$ 0.9	79.5 $\pm$ 1.6	76.5 $\pm$ 0.4
Dialogue Task Segmenter (DTS)	BERT-Base (All*)	110 M	22.5 $\pm$ 0.3	<b>93.4 <math>\pm</math> 0.1</b>	58.7 $\pm$ 0.4	69.6 $\pm$ 0.4
	BERT-Base	110 M	19.1 $\pm$ 0.4	75.8 $\pm$ 0.7	83.6 $\pm$ 0.7	77.5 $\pm$ 0.4
	BERT-Large	340 M	18.4 $\pm$ 0.2	77.0 $\pm$ 1.7	83.6 $\pm$ 2.8	78.1 $\pm$ 0.5
	T5-Base (Enc-only)	110 M	<u>17.7 <math>\pm</math> 0.2</u>	77.9 $\pm$ 0.7	84.2 $\pm$ 0.5	<u>79.0 <math>\pm</math> 0.1</u>
	T5-Base (Enc-Dec)	220 M	18.1 $\pm$ 0.6	77.9 $\pm$ 0.3	82.9 $\pm$ 1.6	78.5 $\pm$ 0.8
	T5-Large (Enc-only)	335 M	18.1 $\pm$ 0.2	77.4 $\pm$ 0.4	84.1 $\pm$ 0.4	78.6 $\pm$ 0.3
	T5-Large (Enc-Dec)	770 M	<u>17.7 <math>\pm</math> 0.2</u>	<u>79.1 <math>\pm</math> 0.8</u>	81.9 $\pm$ 0.9	78.5 $\pm$ 0.2
	T5-3B (Enc-only)	1.5 B	<b>17.0 <math>\pm</math> 0.4</b>	<u>78.3 <math>\pm</math> 1.0</u>	<u>85.9 <math>\pm</math> 0.9</u>	<b>80.0 <math>\pm</math> 0.2</b>

Table 3: Results on the ConvRecipes’s test set from an average of 3 runs per model. *All\** indicates that the model was trained on the set of all recipes crawled, in their original form.

# DTS Model Size Influence

- **Model size** impacts performance.
- Going from BERT-Base to BERT-Large improves performance.
- T5 not conclusive with improvements only in Enc-Dec.
- Largest model T5-3B, achieves the best performance.

	Model	# Params	$P_k \downarrow$	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$
Baselines	Rand <sub>0.5</sub>	-	35.4 $\pm$ 0.3	59.9 $\pm$ 0.5	49.7 $\pm$ 0.8	51.7 $\pm$ 0.6
	Rand <sub>0.75</sub>	-	28.3 $\pm$ 0.5	61.2 $\pm$ 0.4	75.0 $\pm$ 0.9	65.2 $\pm$ 0.6
	Every <sub>1</sub>	-	23.3	60.9	<b>98.8</b>	73.8
	Every <sub>2</sub>	-	37.9	59.6	37.9	44.9
	TextTiling	-	28.4	58.7	67.7	61.4
	CrossSeg	110 M	19.5 $\pm$ 0.4	77.5 $\pm$ 0.9	79.5 $\pm$ 1.6	76.5 $\pm$ 0.4
Dialogue Task Segmenter (DTS)	BERT-Base (All*)	110 M	22.5 $\pm$ 0.3	<b>93.4 <math>\pm</math> 0.1</b>	58.7 $\pm$ 0.4	69.6 $\pm$ 0.4
	BERT-Base	110 M	19.1 $\pm$ 0.4	75.8 $\pm$ 0.7	83.6 $\pm$ 0.7	77.5 $\pm$ 0.4
	BERT-Large	340 M	18.4 $\pm$ 0.2	77.0 $\pm$ 1.7	83.6 $\pm$ 2.8	78.1 $\pm$ 0.5
	T5-Base (Enc-only)	110 M	<u>17.7 <math>\pm</math> 0.2</u>	77.9 $\pm$ 0.7	84.2 $\pm$ 0.5	<u>79.0 <math>\pm</math> 0.1</u>
	T5-Base (Enc-Dec)	220 M	18.1 $\pm$ 0.6	77.9 $\pm$ 0.3	82.9 $\pm$ 1.6	78.5 $\pm$ 0.8
	T5-Large (Enc-only)	335 M	18.1 $\pm$ 0.2	77.4 $\pm$ 0.4	84.1 $\pm$ 0.4	78.6 $\pm$ 0.3
	T5-Large (Enc-Dec)	770 M	<u>17.7 <math>\pm</math> 0.2</u>	<u>79.1 <math>\pm</math> 0.8</u>	81.9 $\pm$ 0.9	78.5 $\pm$ 0.2
	T5-3B (Enc-only)	1.5 B	<b>17.0 <math>\pm</math> 0.4</b>	78.3 $\pm$ 1.0	<u>85.9 <math>\pm</math> 0.9</u>	<b>80.0 <math>\pm</math> 0.2</b>

Table 3: Results on the ConvRecipes’s test set from an average of 3 runs per model. *All\** indicates that the model was trained on the set of all recipes crawled, in their original form.



# Conversational Tasks Statistics

	# Steps	# Tokens	Exact Match	= # Steps	+ # Steps	- # Steps	$\Delta\text{Steps} \leq 1$
Human Annotation	5.86	19.21	-	-	-	-	-
Method	Every <sub>1</sub>	9.29	5.00%	5.33%	94.67%	0.00%	24.00%
	Text Tiling	6.32	7.00%	24.00%	49.33%	26.67%	58.67%
	CrossSeg	6.08	13.33%	30.67%	36.22%	33.11%	68.11%
	DTS T5-3B (Enc-only)	6.48	17.00%	27.56%	46.44%	26.00%	68.44%

- All methods tend to **overestimate** the number of steps.
- **Exact match** is higher in T5-3B model due to its ability to capture the segmentation patterns.
- **Non-trivial balance** so it is important to optimize a **combination of various aspects**.

# User Evaluation (1)

- Compare the original **web-based** with the **predictions** of a model.
  - 6 annotators, 50 recipes, and 3 annotations per each.
  - Annotate which segmentation is the **best** (web vs. model).
  - Annotate each segmentation in a **1-5 scale** considering a conversational setting.

<b>Example 1 - Soy Garlic Steak (Web)</b>	<b>Model Output</b>
<ol style="list-style-type: none"> <li>1. In a small bowl, mix vegetable oil, soy sauce, vinegar, ketchup, and crushed garlic. Place flank steak in a large resealable plastic bag. Pour the marinade over steak. Seal, and marinate in the refrigerator at least 3 hours.</li> <li>2. Preheat grill for high heat.</li> <li>3. Oil the grill grate. Place steaks on the grill, and discard marinade. Cook for 5 minutes per side, or to desired doneness.</li> </ol>	<ol style="list-style-type: none"> <li>1. In a small bowl, mix vegetable oil, soy sauce, vinegar, ketchup, and crushed garlic.</li> <li>2. Place flank steak in a large resealable plastic bag. Pour the marinade over steak. Seal, and marinate in the refrigerator at least 3 hours.</li> <li>3. Preheat grill for high heat.</li> <li>4. Oil the grill grate. Place steaks on the grill, and discard marinade. Cook for 5 minutes per side, or to desired doneness.</li> </ol>

# User Evaluation (2)

- Preference for the **model's segmentation** 86% of the time.
- In conversational suitability the model's predictions achieved a **much higher** score (3.72) than the original recipes (2.63).
- **Takeaways:**
  - Original recipes are **not dialogue suited**.
  - Models are able to increase the suitability of a recipe to a **conversational-friendly** format.

	Web-based	T5-3B (E-only)
Rating 1	18.0%	3.3%
Rating 2	36.0%	12.7%
Rating 3	18.7%	20.7%
Rating 4	20.0%	35.3%
Rating 5	7.3%	28.0%
Best	14.0%	86.0%
Conv. Suitability	2.63	3.72

Table 5: User study results comparing the original web-based segmentations with T5-3B (Enc-only) model predictions. (Conversation Suitability is given on a 1 to 5 scale).



# Conclusions

# Conclusions

- **ConvRecipes Corpus:**

- A dataset with dialog-suited instructional text in the recipe's domain.
- Instructional text found online is not optimal for a conversational setting.

- **Dialogue-Task Structurer (DTS):**

- Models that work at a token-level to capture conversational task segmentation patterns.
- Abstracts less information since each token embedding is contextualized on the full-task.

- **Real-World Improvement:**

- Ability of the model to improve the original recipe 86% of times.
- Better user experience in a conversational-assistance scenario.

# Thanks!

## Grounded Complex Task Segmentation for Conversational Assistants

SIGDIAL 2023 - Prague, Czechia

Rafael Ferreira - [rah.ferreira@campus.fct.unl.pt](mailto:rah.ferreira@campus.fct.unl.pt)

David Semedo - [df.semedo@fct.unl.pt](mailto:df.semedo@fct.unl.pt)

João Magalhães - [jmag@fct.unl.pt](mailto:jmag@fct.unl.pt)



NOVA University of Lisbon, NOVA LINCS, Portugal

**Acknowledgments:** This work has been partially funded by the FCT project NOVA LINCS Ref. UIDP/04516/2020, by the Amazon Science - TaskBot Prize Challenge and the CMU|Portugal projects iFetch CMUP LISBOA01-0247-FEDER-045920), and by the FCT Ph.D. scholarship grant UI/BD/151261/2021.