Did You Get It? A Zero-Shot Approach To Locate Information Transfers In Conversations

Eliot Maës, Hossam Boudraa, Philippe Blache, Leonor Becerra-Bonache

Did You Get It? A Zero-Shot Approach To Locate Information Transfers In Conversations

Eliot Maës[1], **Hossam Boudraa**[1,2], **Philippe Blache**[3], **Leonor Becerra-Bonache**[1]
[1]Aix Marseille Univ, CNRS, LIS, Marseille, France
[2] Department of Computer Science, Faculty of Sciences Dhar El Mahraz,
Sidi Mohamed Ben Abdellah University, Fez, Morocco
[3]Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
{eliot.maes, leonor.becerra}@lis-lab.fr
hossam.boudraa@usmba.ac.ma,
blache@ilcb.fr

Abstract

Interaction theories suggest that the emergence of mutual understanding between speakers in natural conversations depends on the construction of a shared knowledge base (*common ground*), but the details of which information and the circumstances under which it is memorized are not explained by any model. Previous works have looked at metrics derived from Information Theory to quantify the dynamics of information exchanged between participants, but do not provide an efficient way to locate information that will enter the common ground. We propose a new method based on the segmentation of a conversation into themes followed by their summarization. We then obtain the location of information transfers by computing the distance between the theme summary and the different utterances produced by a speaker. We evaluate two Large Language Models (LLMs) on this pipeline, on the French conversational corpus Paco-Cheese. More generally, we explore how the recent developments in the field of LLMs provide us with the means to implement these new methods and more generally support research into questions that usually heavily relies on human annotators.

Keywords: Natural Conversation, Summarization, Information Location, Thematic Segmentation, LLMs

1. Introduction

The ability that humans have to understand one another, despite divergences in experiences and characters, remains difficult to fully explain. Different works have addressed this question by studying the success and more generally the quality of interactions. In these works, communication success relies on one's ability to share information with others in a way that ensures that they are properly understood. Interaction theories link that ability to the progressive alignment of linguistic representations between participants in a conversation (Pickering and Garrod, 2004, 2021). In these frameworks, collaboration on joint tasks as well as competition or debate hinge on the speakers' ability to build a set of shared knowledge, also known as *common ground*. The quality of an interaction would thus be correlated to the ability to build and expand such mutual knowledge, which would be related to the occurrence of a convergence effect at multiple levels between participants (lexical, syntactic, prosodic, gestures, behaviors, etc). There is however no exhaustive description of how and under which circumstances information enters the common ground. One hypothesis is that the convergence phenomenon, and especially its semantic component, is indicative a specific coordination between participants in terms of information exchange, which can be further analyzed by studying the amount of exchanged information and its dynamics during a conversation.

In this perspective, the main question remains to accurately identify the moments where such information exchanges occur in the conversation. This task is usually performed by using dialogue act classifiers (Li et al., 2019) and locating utterances with the "`inform`" label; however even with detailed annotation schemes, the range of this class remains very broad (`inform` constitutes in average 40% of the labels). Previous works have also explored how information transfer could be studied through the prism of Information Theory metrics (Shannon, 1948). In particular, Xu and Reitter (2018) and Giulianelli and Fernández (2021) both explore patterns in the evolution, throughout the conversation, of a metric dubbed entropy and defined as the surprisal of words averaged over an utterance. The finer results of such metrics are however highly dependent on the training of a given model, which makes their use adequate for exploring general patterns, but less so for the exploration of unique or more localised phenomena.

Neither approach constituting a reliable method for accurately locating information transfers in the conversations, we take with this paper a different approach: rather than looking at metrics trying to capture the amount of information throughout the conversation, we turn to the analysis of conversation summaries to try and locate important informa-

tion. Indeed, the goal of the summarization task is to produce a concise overview of a document and of the topics broached. In this case, the annotator (whether human or machine) takes a stand on what information matters and should be reported, and where. This could be linked to the online behavior of humans in a conversation, who instead of strictly memorizing new information rather sort such information depending on their own interests and its importance to the aggregated conversation so far (which might be revised later on). Our idea is then to use summarizing tools to help in locating when (and which) important information is exchanged during a conversation.

As it is usually the case with new NLP questions, the first issue concerns the dataset: summary annotation is expensive to obtain, especially on unannotated corpora that differ from what was previously recorded (either in language or in task). We turn then to Large Language Models (LLMs) for generating the summaries of interest. The recent release of advanced models with emergent abilities that equate the performances of laymen in terms of annotation proficiency on some tasks (Huang et al., 2023) is a revolution, considering annotation is paramount for model training and analysis of phenomena in any data. It has however been found out that those models have a propensity to hallucinate that is not to be neglected, so caution is advised when relying only on their knowledge (Reiss, 2023).

Considering this context, we explore the possibility of using zero-shot models to locate information in conversations, relying on summaries from those models matched to conversation utterance: the more similar the turn to the summary, the more important the information is for the conversation and the stronger the information transfer. We further explore the models' ability to segment a conversation into themes as previous works on information transfer have underlined that interesting phenomenon can be located around theme changes.

Our contributions are manifold: first, we implement new methods to try and find information transfer in conversation. Second, we explore LLM prompting and results on the tasks of theme segmentation and summarization of conversation. We explore metrics to evaluate automated summaries and extract information about utterances despite the possibility that models might hallucinate. Finally, we compare our results to trends in information-sharing obtained by Information Theory metrics. The method we propose is generic enough to be applied regardless of the topic of conversation. It targets directly the question of locating precisely the position of information transfers in conversation, which is paramount both for better understanding the organization of dialogues (studying common ground instantiation) and correlating linguistic behavior with information from other modalities (for instance neuro-physiological).

The paper is organized as follows. We present in Section 2 the state of research in the different fields we take our inspiration from. Section 3 lists the models we used and the dataset on which we applied our methods. Section 4 and 5 respectively detail the different experiments we realised on our corpora, and their results. We conclude on our method in Section 6.

2. Previous Works

2.1. Summarizing Conversations

Automatic text summarization is usually tackled in two different manners, both with their advantages and drawbacks. The extractive method consists in selecting and copying content (part of to full sentences) from the source text and aggregating it into a summary, creating a disjointed but faithful representation of the original document. The abstractive method, on the other hand, is not restricted to rearranging existing segments; it can produce new sentences or use synonyms to describe a document. Summaries created in such a way are closer to those humans would naturally produce; however they suffer from two drawbacks. First, hallucinations are commonly found with such a technique (from simple logical fallacies to links made with completely unrelated content for models trained on larger sets of data). Secondly, the information compression allowed by the rephrasing process makes it more difficult to pinpoint the origin of an information in a summary, as well as the evaluation of the accuracy of summary models.

Few works cover the specific case of natural conversation summaries, as this topic raises several difficulties. First, the relevant information might be distributed across several utterances and mixed with conversational fillers and feedbacks, rendering statistical methods inaccurate. Unlike with written texts, different participants collaborate in a conversation, using their shared knowledge to infer meaning, correcting their thoughts on the fly, etc. This results in utterances not always grammatically well-formed or with explicit (or even clear) meaning. Similar to written sources however, different subjects might be covered throughout a single conversation, some subjects being touched on for just a few utterances, others being brought up several times over the conversation. As a consequence, models context span must be long enough to be able to locate those dependencies. A last point to be raised is the limited training data, with most of the research relying on the few datasets publicly available, such as ICSI (Janin et al., 2003)

and AMI (Kraaij et al., 2005) for English, DECODA (Bechet et al., 2012) for French, which focus more on meetings and task-oriented discussions than natural conversations.

Several studies have tackled these questions, building on BART (Lewis et al., 2019) variations as the model can be fine-tuned for both extractive and abstractive summarization. To compensate the relatively short context span of the model, methods to adequately split the conversation and combine summaries of smaller windows of discussion are proposed (Liu and Chen, 2021; Zhang et al., 2021). Those methods achieve ROUGE scores similar to the levels reached with written text (ROUGE-1 $\simeq$ 53, ROUGE-2 $\simeq$ 22 on AMI for Liu and Chen (2021)). More recently, Zhou et al. (2022) compared the performances of BART-based models to that of generative models such as T5, all models being fine-tuned on conversations thus not leveraging the model zero-shot abilities.

With the emergence of models equipped the ability to produce more complex language, the limitations of the most commonly used metric for evaluating automatically generated summaries, ROUGE (Lin, 2004), have been pointed out. Indeed, as this metric relies on comparing n-grams (1, 2, longest) between original and generated summaries, use of synonyms or rephrasing becomes strongly discouraged during training, reflecting poorly on actual model capabilities. Several alternatives have been proposed, like evaluating sentence embeddings with BERTScore (Zhang et al., 2019) and MOVERScore (Zhao et al., 2019) so as to better take rephrasing into account, or measuring summary factual consistency using natural language inference (NLI) inspired methods with SummaC (Laban et al., 2022). More recently, Liu et al. (2023) proposed a framework for using ChatGPT4 to evaluate summaries, using chain of thought to assess summary coherence within itself and with respect to the source article.

We reckon that those methods relying on similarity measures can also be used in different ways, namely to link information mentioned in the summary back to its location in the source document.

2.2. Utilizing Large Language Models for Annotation

Since obtaining quality annotations for any new dataset usually is a long and costly process, new procedures are regularly proposed for accelerating this step. The emergence of crowdsourcing platforms was a revolution for annotating large batches of new data that algorithms could be trained on; this solution is not applicable to all research questions as some annotation may require expert knowledge. In terms of automated methods, models could be used to generalize based on rules observed from a small set of annotated data. The emergence of LLMs has however further expanded both the use cases and the performances of such algorithms. The release of ChatGPT[1], then LLAMA (Touvron et al., 2023) and other models, has attracted massive public and academic attention. Several studies have discussed the promising zero-shot (prediction without any specific training on a task) applications of such models, which reach, and in some cases even exceeds, the performances of crowdworkers on annotating tasks (Gilardi et al., 2023; Zhu et al., 2023; Kuzman et al., 2023; Huang et al., 2023).

Caution is however advised when relying on such models. Bang et al. (2023) have found out that much like its predecessors, ChatGPT suffers from hallucinations, making logical fallacies when aggregating content or inventing facts to try and give context to content. Furthermore, results consistency and dependency are highly dependent on hyperparameters, in particular temperatures (with lower temperatures being more reliable) as well as variations in prompting (Reiss, 2023).

2.3. Thematic Segmentation and Information Transfers in Conversation

Up until recently, information transfers in language have mostly been studied from the perspective of psycho-linguistic research since the definition of *surprisal* in Hale (2001) and the observation of a particular correlation between reading times and surprisal (Monsalve et al., 2012; Frank et al., 2015). As such, the cognitive load associated to parsing a sentence appears correlated to that sentence predictability and information content. Concurrently, analyses of information transfer at a more global scale emerged, with Qian and Jaeger (2011) then Xu and Reitter (2016) showing a correlation between topic shift and entropy variations in discourse, both written and oral. Indeed, the beginning of a new topic corresponds to the introduction of new information into the context, which causes higher uncertainty in the conversation, especially in the utterances of the speaker who introduces it. Giulianelli and Fernández (2021) later expanded on Xu and Reitter (2018) with deep language models.

Defining a metric that can be used not only for pattern examination (at global or smaller scale) but also for more qualitative analysis of information variations in the conversation, thus locating new behaviors of interest, is however not so simple. An attempt at such a method was realised in Maës et al. (2022), looking at entropy peaks, concluded at the limitations of using metrics directly based on extracted probabilities for tokens to appear. Many

[1] https://chat.openai.com

factors can indeed be involved (model design, fine-tuning, dataset quality, random seed) in affecting the results. For this reason, we offer another angle to try and achieve this goal.

3. Models and Datasets

3.1. Models

In order to develop processes that could be applied to any new corpora, we explored generative models trained (among other languages) on French, with a large context span. We leverage their zero-shot capabilities rather than fine-tuning them on tasks.

ChatGPT ChatGPT is a large language model developped by OpenAI, built upon the innovations and improvements of its predecessors. In terms of training strategies, ChatGPT is a sibling model to InstructGPT, which means it employs instruction and reinforcement learning from human feedback to enhance its overall performance and adaptability. Upon its release, ChatGPT has garnered considerable attention from researchers (Leiter et al., 2023), showing both enthusiasm at the capabilities of the model and its potential uses for data annotation, and carefulness as to the models still partially unknown limits and biases. For all experiments, we used the ChatGPT `gpt-3.5-turbo` with default hyper-parameters saved for the temperature, which was set to vary.

Vigogne Vigogne-13B-Instruct (Huang, 2023) is a fine-tuned version for French instructions of the LLaMA-13B (Touvron et al., 2023) model. Much like ChatGPT, LLaMA is an auto-regressive model designed for dialogue use-cases. It is however advertised as being limited to English; but as weights for the model have been made available, fine-tuned version for other languages have been made available since the model's initial release.

As both models are LLMs and prone to hallucinate during text generation, we query them with varying temperatures and output lengths.

3.2. Dataset

Following previous research on the topic of information transfer, we focus for our analyses on the Paco-Cheese corpus (Amoyal, 2021; LPL, 2018).

Paco-Cheese (PC) is a multimodal corpus containing audio and video recordings of 26 interactions between dyads of participants. Conversations are in French and lasting 15 to 20 minutes; participants were given a short prompt to read to elicit conversation before continuing the talk on the topics of their choice. For 16 out of the 26 recordings, interactions happened between participants that were not acquainted. Manual transcription was obtained and automatically aligned to the audio signal. Consequently, the speech segments we consider here are utterances or *inter-pausal units* (IPUs), segments of speech of which boundaries are defined by pauses longer than 200ms of silence. The corpus is furthermore enriched with annotations for noise, laugh, pauses, feedbacks, head nods and smiles (Amoyal, 2018; Amoyal and Priego-Valverde, 2019). Expert thematic annotation has been added to 16 of the dialogues.

Our reasons for choosing a corpus not previously annotated with summaries are several. First, previous research was done on this corpus on the topic of information transfer. Secondly, few datasets of completely free conversation exist for French, especially in the public domain (the aforementioned DECODA dataset is more task-oriented discussion); and finally, where such conversation exists, annotation including thematic annotation and summaries for the given themes is not found.

3.3. Information Transfer Annotation

As the dataset had not been previously annotated in summaries nor information exchanges, we provided manual annotation on a subset of the dataset, so as to be able to evaluate our methods (and initiate a future resource for supervised learning). 52 segments of dialogue (each segment corresponding to a theme) were annotated for summaries. Out of these, 27 segments were selected and annotated for information content by 4 experts. Annotators were aware of the annotated theme but were instructed to read the dialogue first, so as to be able to better judge of the segment's content, and then rank each IPU in a 3-level scale:

Level	Description
1	The IPU contains *major information*, strongly correlated to the currently discussed theme, which has *never been mentioned* before
2	The IPU contains information, but of *secondary importance* to the conversation; it can include utterances that detail previously mentioned information
3	The IPU contains little to *no information*, or a repetition of aforementioned information

Table 1: Description of classes used in the information transfer annotation

Our definition of the information is then driven by

the task: assessing a level of information taking into account a given theme. No other consideration (for example concerning socio-emotional aspects) is involved. Inter-annotators agreement was computed using `statsmodels`'s implementation of Fleiss' Kappa. Since the "informativity" of an utterance would normally be seen as more of a continuum, annotators had a different understanding of what "information relevant to the conversation" means, resulting in some annotators being more conservative than others and more values being labeled as non-informative ("3"). Overall agreement between annotators is 0.340 (fair agreement) when considering all categories, and 0.454 (moderate agreement) when only considering if an IPU had or did not have information ({1,2} labels vs {3}). Individual dyads had higher agreement. More information on the levels of agreement can be found in Appendix A.

To simplify later analyses, we unify the 4 sets of annotations into two new sets, created with diverging strategies. The first synthetic set (ann_s) uses the probability, for each annotator, of a given label to appear when solving conflicts. This goes more in favor of annotators who annotated more values as containing information. The second set (ann_c) pushes values depending the probability for a given annotator to label an utterance as non-informative, thus favorising the labels chosen by more conservative annotators.

For following analyses, we focus on locating informative *vs.* non-informative segments of the conversation.

4. Experiments

In order to automatically annotate the location of information transfers in the conversation, we decided on a three-step pipeline, with a first step for the discovery of themes in the conversation, a second one using the obtained segments for the generation of summaries, and a final step for the analysis of the summaries. The first two steps use LLM prompting for the results. The reasons for this choice, besides being inspired by the literature, are two-fold: first, it makes it easier to control the interactions between the tasks as well as the length of the generated summaries; secondly, at the time of study, models with a large enough span to take into account the whole conversation had not yet been released.

4.1. Thematic Segmentation

The larger input size of ChatGPT-3.5 being lower than required by the conversations in our dataset, we processed the conversations in parts, each file making 3-5 splits. We prompted the model to produce Thematic Segmentation (given in Appendix B) of each of the dataset conversations. In order to simulate several annotators, we queried the model using 21 temperatures (from 0.0 to 2.0 with a step of 0.1). The temperature parameter controls how consistent the response from the model will be, with 0 being more consistent and 2.0 more random; in our case, using low temperatures ensures that we obtain reliable responses, while including higher temperatures mitigate the eventual failures in response parsing (see Appendix C.1 for examples) and varies the results in annotation. On average, 12.4 $\pm$ 2.7 temperatures were exploitable for a given file, meaning that parsing was successful for all conversation splits with this temperature.

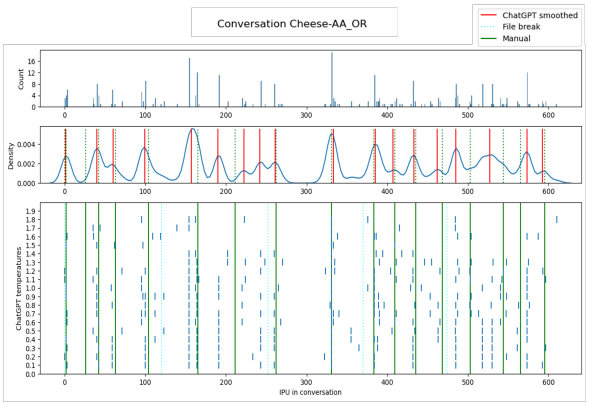

Figure 1: ChatGPT thematic segmentation at different temperatures (y-axis) for one conversation. Locations and frequencies are given as a histogram (top, for all temperatures), a KDE (middle, exploitable temperatures only) and as a raster plot (bottom, for all temperatures). Manual annotation theme breaks are labeled in green. Obtained thematic breaks with ChatGPT are in red.

After parsing the model's answers, we obtained the different themes and the utterance at which they started in the conversation. Considering the slight variation in results between the temperatures, with some temperatures yielding a more sensible segmentation than others, we fitted a gaussian kernel-density estimate (KDE) to the extracted locations for theme starts. We thus extracted a unified theme segmentation by identifying and keeping the n local extrema, where n is the minimum between the number of local extremas, and the average number of themes identified by ChatGPT for the file. The theme label was taken as the most common label among annotators having located a change. An example of this process is given in Figure 1.

Finally, we compared this process to the manual thematic annotation and to thematic segmentation obtained from TextTiling (Hearst, 1997), a statistical method that relies on lexical co-occurrences to compute a similarity score between sentences and segment a text into subtopic shifts. A similar process was attempted using the Vigogne model,

but despite prompt adaptation, we did not obtain any responses that could be exploited.

4.2. Summarizing Conversations

Despite several attempts at prompt engineering, we did not manage to obtain extractive summaries from the model. Such instructions would indeed result in strong hallucinations from the models, despite the witnessed ability of ChatGPT to quote IPUs when segmenting the conversations for themes. We therefore restricted our analysis to the obtained abstractive summaries. Each model was prompted for summaries of different length ("in n words" for ChatGPT, with n taking values in {20,50,100,220}, and "short" and "detailed" for Vigogne as indicating a number of words did not seem to have any effect) at 4 different temperatures. Obtaining several versions of summaries for a given theme of the conversation is useful both to alleviate the issues of hallucinations, and to study whether the information differed between summaries. Simple filters were applied to get rid of summaries with basic issues, such as summary being longer than the input conversation or model yielding summaries in English despite conversation and prompt being in French (details on the amount of summaries filtered out are given in Appendix C.2).

4.3. Evaluating Summaries and Locating Important Information mentioned by a Speaker

In order to link the information mentioned in the summaries back to its location in the conversation, we compare several similarity scores, evaluating the similarity between each sentence s of the summary and each IPU u of the conversation.

The first method we implemented relies on the obtaining the alignment of the sentence embeddings of each (u,s) couple. We use HuggingSpace's implementation of BERTScore[2] and obtain precision, recall and f1 values for such alignment.

Our second similarity metric ($SEmb_{xPOS}$) also relies on embeddings similarity. Using HuggingFace's Feature Extraction Pipeline, we retrieve Vigogne embeddings at the token level for both IPUs u and summary sentences s; we then filter out words using Part-of-Speech tags and stopwords lists (see Appendix D.1 for the detail on PoS tags filtered out), only keeping content words for the distance calculation. Embeddings are then averaged so as to keep one vector for each sentence. Finally, cosine distance is computed on each (u,s) pair.

We took inspiration from SummaC for our final metric and looked to Natural Language Inference models to evaluate whether an utterance can be linked to a summary sentence (and vice-versa). We rely on Hugging-Face implementations of the DeBERTa model[3] (Laurer et al., 2022), fine-tuned on the XNLI dataset for multilingual inference (Conneau et al., 2018) to compute entailment probabilities on our data ($SEnt_{u \to s}$).

We compare the obtained similarities with manual annotation and explore correlations with entropy values previously obtained on this corpora.

Number of themes per conversation	GPT	Manual	Text Tiling
mean	17	16.1	34.4
std	2.2	2.8	8.2
min	13.0	10.0	23.0
max	21.0	21.0	51.0
agreement κ	0.394		0.188

Table 2: Comparison of the number of themes (on average) in conversations according to ChatGPT segmentation, manual annotation, and TextTiling segmentation. Cohen's Kappa agreement score between manual annotation and automated annotation is also given.

5. Results

5.1. Thematic Segmentation

The thematic segmentation obtained from ChatGPT is much closer to the manual segmentation than any out-of-the box algorithm we previously used (see Table 2), both in the number of detected themes and in terms of inter-annotator agreement with the manual annotation ($\simeq$0.4, which according to (McHugh, 2012) falls at the boundaries between fair and moderate agreement). Lower temperatures tend to give slightly better results (details in Appendix E) but as it is impossible to known in advance which temperatures will give the best results, we rely on the unified procedure for a reliable annotation. Distribution of the theme boundaries was stable enough with temperature variation.

Several reasons can be suggested to explain the differences between manual and obtained theme segmentation. Thematic segmentation is a relatively subjective annotation, less in terms of theme content than in terms of setting the exact boundaries where a theme starts and stop. For this reason, the human annotator separated the various themes they annotated with 'transition' moments. Furthermore, the annotation we obtain from all annotators here is only one-dimensional ($theme_1$ |

[2] https://huggingface.co/spaces/evaluate-metric/bertscore

[3] https://huggingface.co/docs/transformers/model_doc/deberta, fine-tuned weights were imported from MoritzLaurer/mDeBERTa-v3-base-mnli-xnli

source	ChatGPT	Vigogne
BERTScore (f1)	0.71 $\pm$ 0.03	0.66 $\pm$ 0.03
$Sconv$	0.51 $\pm$ 0.25	0.67 $\pm$ 0.33
S_{ZS} (entail only)	0.33 $\pm$ 0.10	0.55 $\pm$ 0.24
S_{ZS}	-0.38 $\pm$ 0.13	-0.07 $\pm$ 0.35
Rouge-1	0.25 $\pm$ 0.08	0.13 $\pm$ 0.08
Rouge-2	0.08 $\pm$ 0.04	0.01 $\pm$ 0.02
Rouge-L	0.17 $\pm$ 0.05	0.09 $\pm$ 0.05

Table 3: Evaluation of ChatGPT and Vigogne summaries.

$theme_2$ | ...) and does not truly reflect the complexity of conversations, where subtopics can be nested into broader topics or addressed several times. More accurate models of thematic segmentation, both linguistic (Nakatani and Traum, 1999; Traum and Nakatani, 1999) and computational (Griffiths et al., 2003) have approached the distribution of themes in the conversation as a hierarchical tree. A cutoff value can be used to obtain a more accessible annotation, but there is no information that such a method was used here. This could explain why, sometimes, a human-annotated theme change falls in a local minimum of the unified KDE.

5.2. Summarizing conversations

We establish baseline performances for summarization on the Paco-Cheese corpus, using models not fine-tuned for this particular task. An overview of results is given in Table 3. ROUGE and BERTScore values are average, while SummaC S_{ZS} value is negative, which would imply that summaries actually contradict the more than they are entailed by it (Laban et al. (2022) report scores of $\simeq$ 0.7 for both Summa-C metrics (S_{ZS} and S_{Conv})). Vigogne summaries seem slightly more reliable than those of ChatGPT, but further from the reference.

For ChatGPT, summaries generated using a prompt asking for a smaller number of words performed slightly better than longer summaries, with a more noticeable effect on the ROUGE scores (ROUGE-1 $\simeq$ 34, ROUGE-2 $\simeq$ 11, ROUGE-L $\simeq$ 25 for summaries with a constraint of *20 words*). There was no significant effect of temperatures on the performances of the model. For Vigogne summaries, there were no differences on performances with temperature and length. Despite neither model exactly matching the assigned number of words to use for generating a summary, differences in the distribution of lengths are observable. This effect was especially more prominent with ChatGPT, with only a slight difference in distributions of the number of generated tokens for Vigogne.

5.3. Using Summaries to Locate Important Information given by a Speaker

We finally turn to the prediction of moments of information transfers in the conversation. We consider that the similarities values obtained with the various methods are akin to probabilities of whether an utterance contains information related to a given summary sentence.

Predicting a label based on similarities Since we obtained similarity values for each sentence summary $s \times$ utterance s pair, for various summary length and temperatures, we first need to aggregate the results so as to be able to label each utterance individually. We compare two aggregation strategies. The first one simply takes the max value of a similarity metric for a given utterance. The second method relies on two hypotheses: that the summary will have filtered out the less informative parts of the conversation, and summary sentences for longer summaries will gradually yield lower and lower information, much like the way annotators labelled utterances as more informative and more related to the topic at the beginning of the conversation, than turns that come later. We use a weighted average with decreasing weights to summarize the values from the different sentences for a given temperature and summary length, then take the average value for the utterance over those parameters. We then order the utterances from highest to lowest aggregated similarity and label them based on the class ratios for each annotator. Doing so allows us to obtain better results with individual metrics than using classifiers (LDA, SVC) to predict labels using all similarity metrics. The results obtained with these methods are displayed in Table 4. Detailed results by classes are added in Appendix F. Both strategies perform similarly, but the second method yields a slight improvement in prediction accuracy.

In terms of comparison of the different metrics we defined in Section 4.3, the best results were most often obtained with the $SEmb_{xPOS}$ similarity. Out of the three metrics obtained using BERTScore (precision, recall and f1-score), BERTScore-precision was closest to human judgment and came 2nd compared with other methods. $SEnt_{u \to s}$ came third. Using a voting method between these three metrics did not improve the results.

All methods performed much better than a random baseline (obtained using `scikit-learn`'s DummyClassifier), with stronger improvements for labels which are easiest to annotate (1,3).

In terms of agreement with the synthetic annotators, the metrics matched ann_s (defined as more generous in annotating utterances as containing information) more easily on "informative" classes

annotator	thematic segmentation	labeling	method	subset	κ_3	κ_2
ann_s	manual	annotator ratio	max similarity	all summaries	0.382	0.545
			max similarity	n_words=220	0.377	0.548
			sentence order + max similarity	all summaries	**0.394**	**0.548**
		classification	LDA	5-fold CV		0.408
ann_c	manual	annotator ratio	max similarity	all summaries	0.358	0.489
	ChatGPT		max similarity	all summaries	0.184	0.262

Table 4: Best results obtained in terms of agreement with the synthetic annotators by using the similarity methods.

and ann_c (more conservative) more easily on "non-informative" labels, resulting in a higher agreement score with ann_s than ann_c. The maximum agreement reached with these annotations is slightly under the agreement of these sets with humans ($\kappa_3 \simeq 0.58, \kappa_2 \simeq 0.65$)

We observed no differences in our results depending on whether summaries were generated using ChatGPT or Vigogne, proving that despite differences in summarizing style and issues or hallucinations that might be detected in individual summaries, generating and combining several summaries smoothes out the performances.

Summary length and temperature variation Longer summaries obtained from ChatGPT generation seemed to contain more information to help boost the prediction of non-informative labels in their later sentences, as decreasing the weight given to those decreased the performances of these summaries for classification. Overall, longer summaries performed better for the classification of informative *vs.* non-informative utterances, while middle-sized summaries were more useful to classify relevant ({1}) *vs.* informative but less relevant ({2}) utterances. Combining results from different length in the analysis thus slightly improves performances.

There was no effect of temperature on performance of labeling utterances for information transfers.

Variations in Thematic Segmentation Considering that the annotation of information levels was obtained on the manually segmented themes, we check the viability of our complete pipeline by also obtaining summaries and similarity values for conversation utterances using theme boundaries obtained from ChatGPT. Indeed, it is important to evaluate how changes in the segmentation might affect the prediction of the location of information exchanges, as summaries obtained for themes with different boundies might reflect different moments of a conversation. We observed a clear drop in the accuracy of the prediction of these locations of interest, underlining the dependency to conversation organisation to select informative content.

These results however highlight a dependency to the construction of the summary for results. Since we match utterances of diverse lengths and complexity to summary sentences, shorter utterances will most likely be discarded as being less informative than longer utterances, when that conciseness could be an effect of the conversation process (implied information, etc.). Shorter utterances will also most likely be combined into different propositions in summary sentences, making them less likely to be matched. Experiments with the NLI model used showed that IPUs matching the beginning of a summary sentence rather than its end would more likely be classified as *contradictory* to the summary sentence, which could explains this metric yielded lower performances than others and why the *precision* component in BERTScore performs better than others. Future works might look into discourse simplification tools such as Niklaus et al. (2023) to try and mitigate those effects.

5.4. Comparison to Entropy Metrics

We compared locations of information transfer we obtained to entropy values and peaks locations obtained in previous works. We could not find any correlation between entropy values and manual annotation nor the values of similarity of any measure used; the distribution of entropy values by information content label were almost indistinguishable. Similarly, locations predicted as "informative" by peak values in entropy did not match with annotators' judgments. Those peak locations are fewer in number and mostly fall in the "non-informative" (label {3}) bin. This confirms one of the paper's qualitative estimations that despite seemingly happening close to theme changes, those peak values were often artifacts of the model, which had been fine-tuned on speech corpora but were not most often fell in the label {3} bin.

6. Conclusion

With this paper we explored the possibility to use the zero-shot capabilities of novel models to locate information transfers in natural conversations, using thematic segmentation and summaries on previously non-annotated corpora. This work relied on two main hypotheses.

First, that ChatGPT aligned with human judgement well enough and consistently enough to be used on more peculiar tasks than is usually the case, such as thematic segmentation of natural conversation. We demonstrate that this is indeed the case, enabling us to speed up the annotation of new corpora by relying on automated methods that perform better than previously used, specialized algorithms such as TextTiling that performed difficulty on natural conversation.

Secondly, we hypothesized that summaries could be used as a pointer for information transfer in natural conversations. We established a first definition on how to annotate the question of information transfer in conversation, and provided a small reference for algorithms to be compared to. The different similarity metrics showcased seem to indicate the viability of such a method, though more work is necessary to finesse our method. Future work will also focus on methods to assess and investigate the link between conversation and summary, and summary and utterances, in terms of coverage of information and explainability of model information selection.

Overall, we demonstrate the ability to use LLM pipelines to generate new, unconventional, "subjective" annotations that still correlate strongly with labels that would be obtained from human annotators.

7. Acknowledgements

This work was carried out within the Institute of Convergence ILCB and was supported by grants from France 2030 (ANR-16-CONV-0002), from the CNRS through the MITI interdisciplinary programs through its exploratory research program and from the Institut Carnot Cognition.

8. Bibliographical References

Mary Amoyal. 2018. Analyse du sourire lors des transitions thématiques dans la conversation.

Mary Amoyal and Béatrice Priego-Valverde. 2019. Smiling for negotiating topic transitions in french conversation. In *GESPIN-Gesture and Speech in Interaction*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Frederic Bechet, Benjamin Maza, Nicolas Bigouroux, Thierry Bazillon, Marc El-Beze, Renato De Mori, and Eric Arbillot. 2012. Decoda: a call-centre human-human spoken conversation corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1343–1347.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The erp response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Mario Giulianelli and Raquel Fernández. 2021. Analysing human strategies of information transmission as a function of discourse context. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 647–660.

Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. 2003. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceeding of 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA.

Marti A Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Bofeng Huang. 2023. Vigogne: French instruction-following and chat models. https://github.com/bofenghuang/vigogne.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*.

Taja Kuzman, Nikola Ljubešić, and Igor Mozetič. 2023. Chatgpt: beginning of an end of manual annotation? use case of automatic genre identification. *arXiv preprint arXiv:2303.03953*.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2022. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, pages 1–33.

Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. 2023. Chatgpt: A meta-analysis after 2.5 months. *arXiv preprint arXiv:2302.13795*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Ruizhe Li, Chenghua Lin, Matthew Collinson, and Xiao Liand Guanyi Chen. 2019. A dual-attention hierarchical recurrent neural network for dialogue act classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, et al. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. arxiv abs/2303.16634 (2023).

Zhengyuan Liu and Nancy F Chen. 2021. Dynamic sliding window for meeting summarization. *arXiv preprint arXiv:2108.13629*.

Eliot Maës, Philippe Blache, and Leonor Becerra-Bonache. 2022. Shared knowledge in natural conversations: can entropy metrics shed light on information transfers? In *26th Conference on Computational Natural Language Learning (CoNLL)*, pages 213–227.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Irene Fernandez Monsalve, Stefan L. Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 398–408. Association for Computational Linguistics.

Christine Nakatani and David Traum. 1999. Coding discourse structure in dialogue (version 1.0).

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2023. Discourse-aware text simplification: From complex sentences to linked propositions. *arXiv preprint arXiv:2308.00425*.

Martin Pickering and Simon Garrod. 2021. *Understanding Dialogue*. Cambridge University Press.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Ting Qian and T Florian Jaeger. 2011. Topic shift in efficient discourse production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Michael V Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

David Traum and Christine H Nakatani. 1999. A two-level approach to coding dialogue for discourse structure: activities of the 1998 dri working group on higher-level structures. In *Towards Standards and Tools for Discourse Tagging*.

Yang Xu and David Reitter. 2016. Entropy Converges Between Dialogue Participants: Explanations from an Information-Theoretic Perspective. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 537–546, Berlin, Germany. Association for Computational Linguistics.

Yang Xu and David Reitter. 2018. Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, 170:147–163.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed H Awadallah, Dragomir Radev, and Rui Zhang. 2021. Summˆ n: A multi-stage summarization framework for long input dialogues and documents. *arXiv preprint arXiv:2110.10150*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

Yongxin Zhou, François Portet, and Fabien Ringeval. 2022. Effectiveness of french language models on abstractive dialogue summarization task. *arXiv preprint arXiv:2207.08305*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

9. Language Resource References

Amoyal. 2021. *PACO - Corpus d'interactions dyadiques conversationnelles en Français*. ORTOLANG (Open Resources and TOols for LANGuage). PID https://hdl.handle.net/11403/paco/v1.

LPL. 2018. *Cheese!* ORTOLANG (Open Resources and TOols for LANGuage). PID https://hdl.handle.net/11403/cheese/v2.

A. Human Information Annotation

We computed inter-annotator agreement between all human annotators on our corpus, and with the computed sets of annotation ann_s and ann_c. Detailed values of inter-annotator agreement are given in Table 5. Overall, annotators can be grouped based on how conservative (i.e., how easily they label an utterance as being "informative") they are, with intra-group agreement being high, and inter-group agreement being lower. Synthetic annotators achieve high inter-annotator agreement, as well as overall higher agreement with individual annotators than any other subgroup.

A further analysis of the distribution of labels over the conversation shows that the distribution of the {3} label is uniform over the conversation. For labels 1 and 2, it varies slightly with the annotator but {1} values mostly appear at the beginning of the theme and {2} values later in the conversation (see Figure 2). The distribution is however skewed when comparing length of utterances with respect to their labels, with sentences containing information often being longer than utterances containing no information.

Number of annotators	Agreement	κ_3	κ_2
4		0.340	0.454
3	best subset	0.382	0.498
3	average	0.334 $\pm$ 0.035	0.447 $\pm$ 0.046
2	1rst best	0.493	0.706
2	2nd best	0.383	0.524
2	3rd best	0.327	0.417
2	average	0.326 $\pm$ 0.101	0.439 $\pm$ 0.152
2	ann_s, ann_c	0.792	0.784
2	average $ann_{1..4}$ with ann_c	0.581 $\pm$ 0.13	0.654 $\pm$ 0.15
2	average $ann_{1..4}$ with ann_s	0.597 $\pm$ 0.08	0.657 $\pm$ 0.08

Table 5: Inter-annotators agreement. κ_3 is the agreement with classes defined as described in 3.3. κ_2 only considers agreement on whether or not an IPU contains information ({1,2} labels vs {3}). The first rows focus on human inter-annotator agreement, while the last part considers synthetic sets of annotations (ann_s and ann_c)

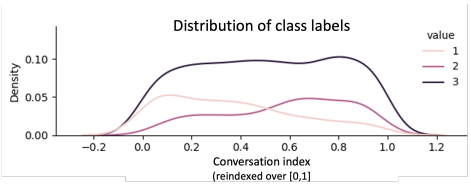

Figure 2: Distribution of class labels over the span of a conversational theme.

B. Model prompts

All prompting was done in the target language, which for this corpus was French. So as to give the model more context and avoid obtaining long descriptions of the situation in the results, the description "this is a conversation between [speaker1] and [speaker2]" was passed into the prompts.

B.1. Thematic Segmentation

> This text is an excerpt of a conversation between AA and OR. When a new speaker takes a turn, their name is indicated between chevron (for instance <AA>). Can you give me the various themes in the conversation and quote the sentence with which they start? for instance: "theme 1: [theme] (introductory sentence: [conversation excerpt]"
>
> conversation: ``` [include the conversation] ```

B.2. Theme Summary

> Sumarise in [n] words the conversation between [speaker1] and [speaker2]: ``` [include the conversation] ```

C. Model Issues

C.1. Thematic Segmentation

On average, out of the 21 temperatures tested (0.0 to 2.0 with a step of 0.1) for thematic segmentation, 12.4 $\pm$ 2.7 were parsed without any issues.

Hereinafter are examples of variations in the response from LLM that resulted (responses 1 to 3) in failures in parsing:

0. Correctly parsed theme

   > thème 0: Organisation des bibliothèques universitaires (phrase d'introduction: "<CG> peut-être par région...si tu as mis ton domicile à Marseille").

1. Simple deviations from the pattern, affecting the parsing from the regular expression (using a different scheme for answering, giving a theme label without associating it to any utterances or vice-versa, etc.)

   > - thème 1: expériences personnelles avec des étudiants en droit(corruption ou pas, introspection quoi') "Après je suis enfin peut-être..."

2. Translations

   Thème 2: Tours/ working wit teams(last twenty phrases bear relevance)

3. Gibberish (can span over 2000 characters, including HTML-like / byte-like segments)

   thème 3: Expérience personnelle de passage de concours (phrase d'introduction: "<cinomen>J avc aila convo gfdpit t'luho vart jnasjbyjh F><CG>Bon là du coup c'est vrai que j'ai pas pensé au autres).

C.2. Theme Summary

Variations and deviations from the expected behavior also appeared in the prompting for summaries. Basic filters were applied to remove summaries that were not deemed usable. Since hallucinations are not easily detectable, those cover responses not made in the target language or taking a form that wasn't adequate for a summary (list-shaped or complete rephrasing of a conversation). As shown in Table 6, those issues mostly arose with ChatGPT responses, despite similar temperatures being used for both models.

Source	ChatGPT	Vigogne
Not French	4,9%	0,1%
Too Long	28,2%	1,6%
List or Fake Dialog	0,2%	0,8%
Total	33,3%	2,5%

Table 6: Percentage of summaries filtered out, out of the 5175 (ChatGPT) and 6716 (Vigogne) generated under all configurations

Despite the observed number of summaries exceeding the threshold in accepted number of tokens, the model overall respected the constraint of lengths with writing summaries, with summaries obtained with the prompt of "50 words" being shorter than those for "100 words" and "220 words".

A qualitative analysis of summaries also showed a curious behavior of the model, which was trying to make sense of the names of the speakers and hallucinating names or descriptions matching the speakers initials.

D. Similarity Metrics

D.1. Filtering out PoS tags for $SEmb_{xPOS}$

We made use of the HuggingFace `transformer` pipeline library to obtain PoS tags and Vigogne embeddings. PoS tags were obtained using a CamemBERT version fine-tuned for classification (`qanastek/pos-french-camembert`). We then excluded from the similarity computation the following list of tags: `'SYM', 'PUNCT', 'YPFOR', 'COCO', 'PREP', 'DET', 'PPER3MS', 'DINTFS', 'PPOBJMS', 'PPOBJMS', 'DETFS', 'DINTMS', 'DETMS', 'PPOBJMS', 'PINDMS', 'XFAMIL'`

E. Thematic segmentation evaluation

A detail account of ChatGPT Thematic Segmentation performances compared to manual annotation and TextTiling is given in Table 7. Lower temperatures overall give better and more consistent results, but the possibility of failures with these temperatures justifies the need to use higher temperatures to get a complete overview of the thematic distribution of the conversation. We obtained our final segmentation based on how unanimously the different temperatures were predicting a theme changed; we observed that while most manually annotated boundaries matched the local maximum of the kernel density estimation function used, some manually annotated also fell in the local *minimum* of the function, meaning all temperatures were in agreement that no boundary were to be defined there.

F. Information Transfer prediction evaluation

While the similarity values obtained for each (s,u) pair were uncorrelated for most metrics (with the exception of the three values returned by BERTScore: precision, recall and f1-score), the predicted output between metrics, computed with Cramer's V, was highly correlated, hence yielding results with very little variation in accuracy across metrics. More information on predictions by label are given in Table 8.

			exact prediction					window=2				
model		nb breaks	TP	precision	recall	f1	kappa	TP	precision	recall	f1	kappa
ChatGPT	mean	187.8	75.4	0.382	0.384	0.388	0.341	71.700	0.362	0.364	0.367	0.326
	std	116.1	48.3	0.115	0.154	0.088	0.118	46.192	0.116	0.149	0.087	0.118
	best (temp=0.4)	302	132	0.437	0.502	0.467	0.441	125.000	0.411	0.475	0.441	0.425
	worst (temp=2.0)	2	0	0.000	0.000		-0.012	0.000	0.000	0.000		-0.024
	unified	287	116	0.404	0.441	0.422	0.394	116.000	0.380	0.445	0.410	0.395
	local min+max	600	137	0.228	0.521	0.317	0.273	144.000	0.228	0.548	0.322	0.256
	all local max	308	119	0.386	0.452	0.417	0.388	121.000	0.370	0.460	0.410	0.392
texttiling		564	93	0.165	0.354	0.225	0.188	96.000	0.170	0.365	0.232	0.179

Table 7: Comparing ChatGPT Thematic Segmentation to manually annotated theme boundaries. TP indicates the number of elements that either directly match a manual annotation or fall within a small window of that point.

annotator	thematic segmentation	class	method	subset	f1-score	baseline f1-score	difference	κ_3	f1-score	κ_2
ann_s	manual	1	max similarity		0.501	0.195	0.306	0.382	0.736	0.545
		2			0.357	0.224	0.132			
		3			0.809	0.575	0.234		0.838	
		1	max similarity	n_words=220	0.476		0.281	0.377	0.738	0.548
		2			0.356		0.132			
		3			0.838		0.263		0.838	
		1	sentence order +max similarity		0.518		0.323	0.394	0.739	0.548
		2			0.369		0.145			
		3			0.853		0.278		0.853	
ann_c	manual	1	max similarity		0.489	0.192	0.297	0.358		0.489
		2			0.246	0.160	0.086			
		3			0.838	0.680	0.158			
	ChatGPT	1	max similarity		0.360		0.204	0.184	0.723	0.262
		2			0.230		0.116			
		3			0.723		0.057		0.522	

Table 8: Best results obtained by the similarity methods in terms of agreement with each annotator. broken down to accuracy to align on each class label. We use `scikit-learn` DummyClassifier to establish baseline performances on this task.