Weakly supervised discourse segmentation for multiparty oral conversations

Lila Gravellier,[1] **Julie Hunter,**[2] **Philippe Muller,**[1,3] **Thomas Pellegrini,**[1,3] **Isabelle Ferrané**[1]
[1]IRIT, University of Toulouse, [2]LINAGORA Labs,
[3]Artificial and Natural Intelligence Toulouse Institute (ANITI)

firstname.lastname@irit.fr, jhunter@linagora.com

Abstract

Discourse segmentation, the first step of discourse analysis, has been shown to improve results for text summarization, translation and other NLP tasks. While segmentation models for written text tend to perform well, they are not directly applicable to spontaneous, oral conversation, which has linguistic features foreign to written text. Segmentation is less studied for this type of language, where annotated data is scarce, and existing corpora more heterogeneous. We develop a weak supervision approach to adapt, using minimal annotation, a state of the art discourse segmenter trained on written text to French conversation transcripts. Supervision is given by a latent model bootstrapped by manually defined heuristic rules that use linguistic and acoustic information. The resulting model improves the original segmenter, especially in contexts where information on speaker turns is lacking or noisy, gaining up to 13% in F-score. Evaluation is performed on data like those used to define our heuristic rules, but also on transcripts from two other corpora.

1 Introduction

Discourse analysis, focusing on pragmatic aspects of text interpretation, especially aspects ranging beyond the level of the sentence, is a long standing domain of Natural Language Processing (NLP), of growing importance for many tasks in NLP such as Machine Translation (MT) (Chen et al., 2020) and summarization (Louis et al., 2010; Xu et al., 2020).

The first level of discourse analysis consists in segmenting a discourse into basic units, which generally correspond to roughly clause-level units of text whose contents provide the arguments to discourse relations such as Explanation, Elaboration, and Contrast (Mann and Thompson, 1987; Prasad et al., 2008; Asher and Lascarides, 2003). Taking the resulting *discourse segments* as input for tasks such as summarization/sentence compression can improve performance over sentence-based approaches (Li et al., 2020; Sporleder and Lapata, 2005; Xu et al., 2020). Segmentation is also crucial for downstream tasks that exploit discourse relations (Chen et al., 2020; Louis et al., 2010; Xu et al., 2020).

Because automatic segmentation on text-based corpora tends to yield good results (above 90% F-score on segment boundary recognition, depending on the language, cf. Zeldes et al., 2019), the segmentation task is often neglected in favor of the prediction of discourse relations. State of the art discourse segmenters, however, generally take sentence boundaries as given, and often benefit from other forms of punctuation, as well as the well-formed sentences found in text-based corpora.

Such systems provide a less stable foundation for discourse parsing of spoken conversations, which is necessary to improve, e.g., real time recommendations from voice assistants and meeting summarization, and to develop more advanced assistants including robots/cobots with conversational capabilities. In these cases, we need to work directly on the audio signal or transcript, and cannot assume sentence boundaries or punctuation (at least not human-corrected punctuation). And the utterances might be far less well formed, as in the following example (translated from our data, see Section 4) in which a speaker interrupts his own thought to ask a question about the name of a town—a question that he ends up answering himself with "no".

(1) so then i did uh we went through uh through uh well uh benghazi is that what it's called that little town there no uh...

Segmentation of such data becomes a more complicated task that requires fine-tuning to discursive properties particular to spoken conversation as well as reliance on acoustic features.

Annotated conversation data is relatively scarce, however, especially for languages other than En-

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1381–1392
November 7–11, 2021.

glish, and existing corpora span a range of different contexts: meetings, written chat, user/system interactions, etc. To address data scarcity, we introduce a method to transfer the knowledge of a segmenter trained on well-prepared, written text to an oral context, with specific supervision given by a latent model bootstrapped by manually defined heuristic rules, as in (Ratner et al., 2020). The rules exploit lexical, syntactic and acoustic features to help identify segment boundaries. This should provide a more general framework than oral segmentation efforts dedicated to specific tasks like MT (Iranzo-Sánchez et al., 2020) or than lexically-based unsupervised approaches (Galley et al., 2003). The contributions of this paper are thus:[1]

- a method to transfer a supervised model based on prepared, written text to spontaneous, multiparty oral conversation using multimodal features and little manual annotation;
- evaluation of the new segmentation model with different input information, on in-domain and out-of-domain data sets, showing large improvements on the written text segmenter;
- a corpus of 7.5 hours of multiparty, spoken conversation in French with gold transcripts, and manual segmentation annotations on (small) development and test sets.

2 Related work

The first automated attempts at predicting discourse structure and discourse segmentation with a rule-based approach can be attributed to (Marcu, 2000). Until recently, discourse segmentation had been generally ignored in discourse parsing, although there were some studies based mostly on statistical models using lexical/syntactic features (Soricut and Marcu, 2003; Fisher and Roark, 2007; Hernault et al., 2010; Joty et al., 2015) and the rule-based approach of (Tofiloski et al., 2009). These approaches are restricted to sentences, however, thus assuming that sentence boundaries are given.

Recently, interest in discourse segmentation was renewed with neural-network sequential classification using contextual embeddings (Wang et al., 2018; Lukasik et al., 2020), though still at the sentence level.

The shared task at the Disrpt 2019 workshop introduced a more general evaluation framework, with multilingual data and segmentation at the level of full texts, with a subtask that did not assume sentence boundaries (Zeldes et al., 2019). The best system, which also used a sequential model over contextual embeddings (Muller et al., 2019), showed the best performance both with and without sentence boundary information.

In work on oral conversation, segmentation has often been approached not as a discourse problem, but as a problem of recognizing "sentences" in order to predict punctuation marks with ngram models and audio features, either to enrich automatic speech transcripts (Batista et al., 2012) or to improve MT (Fügen et al., 2007; Zhang and Zhang, 2020; Wang et al., 2019). Arguably the relevant units for MT are what (Fügen et al., 2007) vaguely call "semantic boundaries". These units are less fine-grained than the segments needed for discourse analysis, however.

(Ang et al., 2005) is one of the first works to simultaneously address dialogue act (DA) segmentation and classification for speech in multiparty meetings (where for the purposes of this paper, we can consider a *dialogue act* as just a dialogue specific term for a discourse segment). They found that a simple prosodic model aided performance over lexical information alone, especially for segmentation. They used pause information for segmentation, and added duration, pitch, energy and spectral tilt features for classification.

(Quarteroni et al., 2011) used conditional random fields to simultaneously segment and label conversations, trained on the Switchboard corpus, with a per-token classification accuracy around 70% (a token is either a segment boundary or not).

More recent approaches make use of neural networks. (Zhao and Kawahara, 2018) proposed a joint segmentation-tagging model using bi-directional Long Short Term Memory layers, in the form of a word sequence tagger for segmentation and a sentence classifier for dialogue act tagging. (Dang et al., 2020) proposed an end-to-end speech-to-dialogue-act recognizer in the form of a single attention-based neural network trained to perform word-level ASR and fine-tuned to perform DA segmentation and classification. Both previous studies use the Switchboard corpus, which, by

[1]The code and a reproduction notebook are available at https://github.com/linto-project/linto-dialogue-act-segmentation, along with the manual transcriptions of the conversations and the dev/test set manual segmentations.

discourse corpora standards, provides a large annotated dataset (1k conversations, 200k dialogue acts, 1.3M tokens) for English. Evaluation measures in speech-oriented work focus either on words (belonging to the right segment or not) or exact segment boundaries of varying types that differ from conventions for written-text segmentation, making direct comparisons difficult.

Our approach is based on the data-programming paradigm (Ratner et al., 2016), a weak supervision framework that has been applied mainly to information extraction problems in NLP, but also recently to discourse analysis (Badene et al., 2019), specifically for discourse structure prediction. We are unaware of similar work on discourse segmentation or on multi-modal text/speech classification problems. The novelty of the data-programming approach is that it requires only a fraction of the data to be manually annotated, for designing heuristics and for evaluation, and is arguably easier to adapt to new data.

3 The data-programming approach

The data-programming approach (Ratner et al., 2016, 2020) can be decomposed into three steps:

(1) LABELING FUNCTIONS: by studying only a small (possibly annotated) development set, experts design a set of heuristic rules or *labeling functions* (LFs) that will be used to automatically label new data. LFs can exploit heterogeneous information sources: other heuristic rules, external knowledge sources, models trained on a similar problem with different data or a different domain, etc. The LFs, which produce "noisy" annotations, need not apply labels to all data points; that is, they may *abstain*.

(2) LABEL MODEL: the LFs are applied to a new data set for which we have no annotation. The predictions of each LF, represented as a label matrix, are then used to train a model of the joint distribution of the accuracy of the different LFs and the (unseen) true labels, based on the LFs' agreement and disagreements on the instances they label. This is similar to majority voting, with LFs also being weighted by their estimated accuracies.

Step (2) is the crucial part of the approach. For the full formal description, we refer to (Ratner et al., 2020, pp 6-7), but in sum: for each LF λ_j and each instance x_i, define the label matrix as $\Lambda = [\lambda_j(x_i)]$, and the vector of the unknown true labels for x_i as $Y = [y_i]$. The label model is an estimate $p_w(\Lambda, Y)$ of their joint probability. It depends on two factors, the label propensity (Lab) of an LF, i.e. how often it assigns a label, and its accuracy (acc)[2]:

$\phi^{Lab}_{i,j}(\Lambda, Y) = \mathbb{1}\{\Lambda_{i,j} \neq 0\}$
$\phi^{acc}_{i,j}(\Lambda, Y) = \mathbb{1}\{\Lambda_{i,j} = y_i\}$

The model is then a log-linear model on the concatenated factor ϕ^i for each instance:

$$p_w(\Lambda, Y) \propto \exp(\sum_i w^T \phi^i(\Lambda, y_i))$$

This is learned without access to Y by minimizing the negative marginal log likelihood given the observed Λ:

$$\hat{w} = \underset{w}{\operatorname{argmin}}(-\log \sum_Y p_w(\Lambda, Y))$$

The predictions $p_{\hat{w}}(Y|\Lambda)$ can then serve as probabilistic labels, yielding a labeled "train" set.

(3) FINAL MODEL: The train set labeled by the label model in step (2) is used to train a supervised model appropriate for the task. If the model does not accommodate probabilistic supervision, one can apply a threshold on the positive class to produce hard labels. An unseen test set is usually annotated manually for evaluation (see Section 4).

We use the snorkel library,[3] which provides implementations of the various steps of data programming, e.g. a framework to develop and evaluate LFs and to train a label model.

While manually designing LFs still requires some human effort and expertise, it has proven to be less labor-intensive and more reliable than massive data annotation on some tasks (Ratner et al., 2020, user study, pp 16-17) and has other benefits: it does not depend on the size of a data set, it is arguably easier to adapt to a different context (and to some extent a different language), and, as we show below, does not require the final supervised model to have access to all of the information initially exploited by LFs. In our case, this means that the annotations are generated by LFs that exploit both textual and acoustic features, but the final model is only trained on transcribed text.

4 Data

Our research aims to improve the conversational capacities of automated assistants in French. As such, we develop our segmentation model using

[2] The full model can learn dependencies between LFs; we opted to ignore these based on results on the development set.

[3] snorkel.org

a new corpus of around 7.5 hours (7hr, 40min) of multiparty conversations in French.

The interactions come from real meetings recorded in an industrial setting and can be divided into three types: (i) presentations: one person presents their work and other participants (7-12 participants (p)) ask questions during or after the presentation; (ii) weekly update meetings (4-5p): participants present their work from the previous week to a manager, field questions, and describe their next steps; and (iii) informal discussions (4-6p): about work related topics or more personal topics, such as vacation plans. The corpus contains a total of 11 meetings ranging in length from 8.5 minutes (type (iii)) to 84 minutes (type (i)).

The entire corpus was transcribed, first automatically and then corrected by hand, so as to include any word that can be heard on the transcript. Speaker turns—maximal, continuous sequences of words that can be attributed to a single speaker——were also manually corrected. Transcripts were aligned with the audio files using JTrans (Cerisara et al., 2009) to produce time stamps for each word.

The interactions, as is typical of meetings and informal conversations, were not prepared in advance, leading to disfluencies, such as hesitations (*uh*, *euh*, etc.), repetitions, self corrections ("j'ai fait euh on est" ≈ *I did uh we went*), and incomplete sentences, as well as linguistic tics and overlapping speech, when people speak at the same time. All of these elements, apart from overlapping speech, are represented in the transcripts. Areas of overlapping speech, which complicate the alignment between the transcript and audio files, and which would have not been captured by an ASR system, were deleted. This led to a removal of 26 minutes, reducing the corpus to 434 minutes (7hr, 14min).

The development set used to design our LFs (see Section 3) contains one weekly update meeting (4p) and one informal discussion about equipment to buy for an office (4p). As shown in Table 1, it contains 9,572 words, or 13.5% of the corpus (without overlapping speech); manual segmentation (to aid in LF design) yielded 1140 segments.

The test set used to validate the whole process contains one presentation (7p) and one informal conversation about vacation (5p), making up 12.2% of the corpus. Manual segmentation yielded around 1100 segments. While some of the participants who asked questions in the test set appeared in other recordings from the corpus, the principal speakers—i.e., the presenter for the presentation and the two people discussing their vacation plans—did not.

Set	N Words	N speakers	Duration
Train	53,692	21	319 min
Dev	9,752	8	59 min
Test	8,833	13	56 min

Table 1: Corpus statistics for each split.

The remaining interactions from the corpus were set aside to be used to train the label model (step 2 in the data-programming approach) and then the final supervised model (step 3). The transcripts for the training set were not manually annotated, as the point of the data-programming approach is to be able to predict them automatically based on weights that the label model assigns to the LFs.

5 Heuristics (Labeling Functions)

5.1 LF design

Segmentation is framed as a token classification problem in our model: for each token of a meeting transcript, a labeling function (LF) may label it as either the beginning of a segment (BOS) or as not being the beginning (NO). The LF may also fail to produce a label, i.e. it may abstain. The LFs, designed based on the development set, exploit a variety of sources of weak supervision, including predictions of supervised models trained on text data, linguistic features of the transcripts and acoustic features of the recordings, as well as potentially interesting mixtures of information sources.

The main source of information in our model is the output of the supervised segmentation model ToNy (Muller et al., 2019), trained on French written data, which we thus try to transfer to oral data within the weak-supervision framework. The corresponding LF, which predicts a BOS label if ToNy predicts a BOS label, can be seen in the top-left box of Figure 1, which summarizes the process of creating new annotations from LFs.

We also used as input an in-house supervised punctuation model trained on transcripts of French conversational and read data, based on logistic regression (cf. Batista et al., 2012), as we assume punctuation of speech transcripts is correlated to discourse segmentation (Quarteroni et al., 2011; Zhao and Kawahara, 2018). The model tags punc-

tuation at word level, using word ngrams around the target word and part-of-speech tags, as well as three parameters extracted from the audio interval that is automatically aligned to the target word: the root mean squared energy orientation of the interval (ascending, descending, stable), the global pitch orientation (ascending, descending, stable, unknown) and the preceding pause duration. Scores are produced at word-level for three classes: period, comma and no punctuation. Overall accuracy on one manually annotated meeting was 53%. As periods were more accurately predicted than commas, and were a more reliable indicator of discourse segment boundaries, our LFs only exploit period probability.

To construct further LFs, we isolated multiple sources of confusion due to typical oral phenomena not present in the training datasets for the models described above, including disfluencies and different uses of discourse markers, which are a very important source of information for ToNy. We also combined the acoustic features described above with other linguistic regularities observed in the development set. Finally, exploiting the SpaCy syntactic parser for French,[4] we looked for correlations between selected grammatical features and gold (manual) BOS labels: (a) part-of-speech trigrams centered at the token where the decision was to be made, (b) syntactic dependency relations between a target token and its syntactic head.

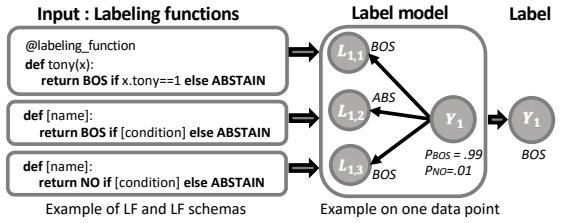

Figure 1: Label model pipeline: generation of labels from "denoised" heuristics, to create training data. The top left box shows an example LF; the two below it show LF schemas. BOS = beginning of segment; NO = no segment boundary at this token.

5.2 LF evaluation

Single LF evaluations can be done on the development set, and include the following factors: (a) *propensity* (*coverage*) of the LF, i.e. the number of instances for which an LF produces a decision (not abstains); (b) *precision*, i.e. the number of correct decisions it makes; (c) *overlaps* with other LFs, i.e. the proportion of instances it covers that are covered by at least one other LF; and (d) *conflicts*, i.e. the proportion of instances it covers for which another LF produced a different decision.

Segmentation of our data is a quite imbalanced binary classification problem: only ≈10% of tokens mark the beginning of a segment. An LF produces at most one possible label (BOS/positive or NO/negative), so a "negative" LF will necessarily have good precision if it has good coverage. Since the label model relies on agreements and conflicts between LFs to assess their reliability and the probability of the true label, it is important to have as many LFs as possible to fire on as many instances as possible. Table 2 shows a few example LFs and their corresponding statistics.

LF	Pol.	Cov.	Ovlp.	Acc
tony	1	0.11	0.09	0.75
period_pred	1	0.06	0.06	0.71
syntax_ngrams	1	0.02	0.02	0.74
stop_pos_type	0	0.33	0.10	0.96
no_disfluency	0	0.24	0.12	0.92

Table 2: Evaluation of a subset of LFs on the dev set. Pol=polarity, as LFs only predict a segment boundary (1) or its absence (0). Cov=coverage of instances. Ovlp=overlap with other LFs. Acc=accuracy of the LF on covered instances of the dev set (≈ precision). LFs shown exploit the following information: label predicted by the written segmenter, ToNy; period predicted at the target token; ngram of dependency types around the token; pos tags typically associated with NO labels in the dev set; disfluencies correlated with "no boundary" decisions. These example LFs are selected by coverage, with a threshold on accuracy at 0.7. The full set of LFs is described in the Appendix.

6 Experiments

We adopt the architecture of (Muller et al., 2019) for our final model (step 3), as this is currently the best discourse segmenter that does not require sentence preprocessing (and the only one with a French model). Using the same architecture also allows us to estimate the degradation of performance from written text to oral speech transcripts.

Experiments were conducted on both the test set from our corpus and transcripts from two other spoken French corpora. An additional advantage of the weakly supervised approach followed here is that

[4] https://spacy.io/models/fr.

our final model can take as input plain transcripts; that is, it does not require all of the information used as input to the label model (pitch, energy, pos-tagging, syntactic analysis, etc.).

For evaluation, we follow the procedure chosen by the Disrpt19 shared task, i.e. we measure F-score on segment boundary detection (BOS tags).

6.1 Label model (step 2) performance

As our final model architecture (based on Muller et al., 2019) does not take probabilistic labels, we discretized the distribution produced by the label model, taking 0.7 as a threshold for positive boundaries based on observation of the development set.

As a first indicator of performance, we evaluated the discretized label model's accuracy on the development set, though we note that because the set was used to design the LFs, it is likely to overestimate model reliability. The label model from step (2) yielded much better accuracy than a simple majority vote between overlapping LFs: 90.8% versus 74%. The F-score on segment boundary prediction was reasonable, but obviously much lower than results on written text: 73% versus 92% in Disrpt 2019, on French text without sentence boundaries.

6.2 Final model setup (step 3)

We trained the final, supervised model with annotations produced by the label model. We borrowed the architecture from (Muller et al., 2019), which is essentially a BERT architecture fine-tuned for sequence tagging, with a Bi-LSTM on top. Then, without changing any hyperparameters from the original setting, we trained it under two conditions, yielding the following three evaluation conditions:

1. ToNy$_W$: our baseline; the model of (Muller et al., 2019), trained solely on written text.
2. ToNy$_{W+O}$: the result of fine-tuning ToNy$_W$ using annotations on our corpus predicted by the label model.
3. ToNy$_O$: a model trained from scratch on the annotations predicted by the label model.

We used the implementation made available by (Muller et al., 2019) for the Disrpt19 task for plain text segmentation.[5] For the baseline and the fine-tuned models, we used the corresponding French model published online.[6]

[5] https://gitlab.inria.fr/andiamo/tony
[6] https://zenodo.org/record/4235850

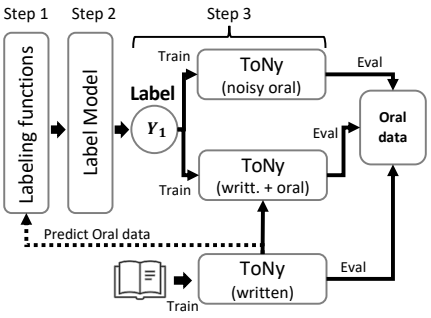

Figure 2: Training and evaluation configurations.

6.3 In-domain data set configurations

In-domain evaluation was performed on our test set, which contains two conversations, with a total of 8,833 tokens and ≈1100 gold segment boundaries.

As we aim to develop a robust model of segmentation for oral conversation to improve models of real-time recommendations or meeting summarization, we cannot always assume perfect transcripts and speaker turn identification. Part of our goal was thus to see if our LFs capture enough information to reliably segment transcripts in the presence of automatic speech recognition (ASR) and speaker change detection errors. We thus performed our evaluations on multiple data conditions.

First, taking the corrected transcript as given (without punctuation), we evaluated our model using (1) gold (manually tagged) speaker turns, (2) automatically predicted turns, and (3) no turn information (just continuous text).

Second, we evaluated our model on the output of the ASR system LinSTT,[7] as it is an open source system for French. In this configuration, only words, not speaker turns, are predicted.

Predictions for speaker change at word level, using Pyannote.audio[8] (Bredin et al., 2020), was rather low on our dataset, with 0.38 precision and 0.15 recall on finding speaker changes, so it serves as a good robustness test. LinSTT's performance on our corpus was also rather low, with a ≈43% word error rate, though this is unsurprising given the difficulty of transcribing conversational speech.

6.4 Out-of-domain data sets

Final evaluations were also performed on transcripts from two out-of-domain datasets: the Rhap-

[7] https://github.com/linto-ai/linto-platform-stt-standalone-worker
[8] https://github.com/pyannote/pyannote-audio.

sodie treebank[9] and the ESTER V2 corpus of broadcast news speech (Galliano et al., 2009).

The Rhapsodie treebank is a syntactically annotated conversation corpus; we used the version available via the Universal Dependency corpus, under the reference "Spoken French".[10] The gold segmentation for Rhapsodie is syntax-oriented and meant to be close to sentences, so not necessarily in line with our definition. ESTER V2 is a French broadcast news dataset built in 2008-2009 for ASR system benchmarking.[11]

The test set for Rhapsodie has 9,850 tokens and 730 segment boundaries. For ESTER V2, we used a single recording comprised of 20 minutes of speech from 26 different speakers, with about 3700 tokens and 354 segment boundaries.[12] Broadcast news speech is mostly prepared speech, closer to read speech and written text data than to the conversational data in our main corpus or Rhapsodie. Evaluation on ESTER thus sheds light on how our model performs on data that fall between fully oral spontaneous speech data and text data.

Train turns	Test turns Cfg	gold	detected	no
gold	O	73.6	57.6	56.3
	W + O	**73.7**	**58.3**	**56.9**
detected	O	66.5	62.0	60.7
	W + O	*69.4*	**63.4**	**60.9**
no	O	60.5	62.5	**64.2**
	W + O	*62.9*	**63.6**	**64.2**
Baseline	W	71.6	53.6	51.1

Table 3: Evaluation of the final models (O and W+O) trained on the noisy annotations from our label model, according to the configuration of the final model and the type of transcript input used for train and test. Test results for the baseline (W), (Muller et al., 2019)'s French model, are not dependent on the oral train set. The best configuration for each train/test speech turn origin is in bold if it beats the baseline for the same testing condition, in italics if not.

7 Results and discussion

7.1 Comparing training conditions

Table 3 summarizes the main experiment, where we compare ToNy$_W$ (W), the baseline segmenter trained on written text; ToNy$_O$ (O), the same architecture trained from scratch on our conversation data; and ToNy$_{W+O}$ (W+O), the architecture fine-tuned on the conversation data. The different train/test inputs are labelled as 'gold': gold speaker turns, 'det': automatically detected turns, and 'no': no speaker turn information.

We can see that the fine-tuned model (W+O) outperforms the model trained from scratch (O) in all configurations except when there is no speaker turn information either for training or evaluation, in which case they both attain an F-score of 64.2%.

Our systems beat the baseline in all configurations with a wide margin—from +2 to +10 points—unless gold turns are not used during training but are given at test time (detected/gold and no/gold, center-left and bottom-left regions of Table 3).

Using gold turns during training yields the best results when gold turns are also given at test time (gold/gold); otherwise, scores fall at least 15 points (first three rows of Table 3).

These results might suggest that providing turns at training time prevents our model from learning important features that distinguish points of speaker change from turn-internal segment boundaries. The results for the configurations with no turns during training do not support this hypothesis; however, we note that in the "no turn information" condition, transcripts were cut arbitrarily to fit constraints on input length imposed by BERT (and ToNy), potentially generating random errors, and the LFs were not designed with this in mind.

The above results were obtained with manual transcripts. Using the transcripts produced automatically with the LinSTT system (about 40% WER), the result for the W+O model with no speaker turn information drops from 64.2% to 49.6%, though it still beats ToNy$_W$'s result of 41%.

7.2 Qualitative error analysis

While we do not provide here a quantitative error analysis of our results, a detailed qualitative analysis of errors predicted on the development set was necessary in order to produce and tweak our LFs. We found that one of the most problematic sources of error was the conjunction *et* (and), which is well known to cause segmentation errors due to its dual function as a propositional and nom-

[9] https://rhapsodie.modyco.fr/.

[10] https://universaldependencies.org/treebanks/fr_spoken/index.html

[11] https://catalogue.elra.info/en-us/repository/browse/ELRA-S0338/

[12] File id: 20071218_1900_1920_inter

inal conjunction. The relative pronoun *que* (that) also escaped our LFs at times and, like *et*, led to over-segmentation errors. Both words are highly reliable indicators of segment boundaries and it was difficult to circumscribe the exceptions. Likewise, isolating frame adverbial uses of adverbs such as *aujourd'hui* (today), which often introduce new segments, can depend heavily on intonation and other acoustic information that we did not have time to study in detail and so could not capture with LFs.

A further hurdle that we encountered was in reliably predicting segmentation boundaries that correspond to speaker changes, leading to errors when gold speaker turns were not provided. A more general problem was that for spontaneous conversation, where speakers do not necessarily produce grammatical or even complete utterances, it can sometimes be difficult to say which is the "correct" segmentation. In these cases, the gold segmentation is arguably arbitrary, meaning that disagreements between the gold and the predicted segment boundaries do not tell us much.

Finally, some of the errors that we found were superficial. As noted above, we sometimes had to cut transcripts arbitrarily to respect word limits imposed by BERT, and these cuts were treated as segment boundaries by default. There were also superficial errors linked to the typical French filler word *euh* (um) in which the gold put a BOS on one side of the *euh* but our segmenter put it on the other side, meaning that the segmenter actually predicted the right place to segment the real content of the transcript.

7.3 Ablation study

We tested the impact of audio-related information on the model by removing LFs involving audio features. Table 4 summarizes the results, separated again by the kind of speaker information assumed: gold speaker turns, detected, or no turn information, either at training or test time. Here we focus on the best model from the previous experiments (W+O).

While removing audio-based LFs does not change our fundamental result—our fine-tuned model still outperforms the baseline, at least when gold turns are not given at test time—the results in Table 4 are mixed. Audio LFs clearly improve our scores only in the gold/gold and detected/gold configurations (+2 points), and in the former, removing the LFs causes the model to fall slightly below the relevant baseline score (71.6). The rest of the evaluation shows either no significant difference without audio-based LFs, or even an improvement.

Train turns	Test turns / LFs	gold	detected	no
gold	all	73.75	58.32	56.91
	wo audio	71.49	60.76	60.28
detected	all	69.45	63.40	60.93
	wo audio	67.18	63.25	62.44
no	all	62.93	63.62	64.21
	wo audio	64.58	63.82	64.34

Table 4: Evaluation of the impact of audio-related LFs. Here we show results only for the fine-tuned model (W+O), which had the best scores in the preceding experiment (see Table 3).

We note that in addition to potential errors induced by arbitrary cuts in the transcripts imposed by BERT, our audio rules require perfect alignments between tokens and time stamps. As alignment was done automatically without correction on our train set, this could be another source of error, though a careful study would be required to see.

7.4 Out-of-domain evaluation

We tested the robustness of the models by applying them to transcripts from two other corpora, without additional fine-tuning. We opted to evaluate the model trained with no speaker turn information, as it assumes the least about the target data. The main results are presented in Table 5.

We can see that ToNy$_W$ provides the best score for ESTER, which is broadcast speech and thus closer to prepared or written text than the kind of conversation for which our LFs were designed.

By contrast, our models clearly outperform ToNy$_W$ on Rhapsodie, with ToNy$_{W+O}$ providing the best result, showing a nearly 10 point improvement over the segmenter baseline, ToNy$_W$. All three segmentation models outperformed a second baseline (Baseline-2), a segmentation approach using syntax-based sentence splitting, with ToNy$_{W+O}$ showing a more than 30 point improvement (as evaluated during the ConLL 2018 shared task[13]). When we compare these results to those of

[13]See https://universaldependencies.org/conll18/ in Results, then Sentence information, table fr_spoken.

Table 3 for no speaker turns in training or testing—the setup we use here—we see that W+O loses about 8 points when switching to Rhapsodie.

Corpus	Config	F-score
Rhapsodie	Baseline-2	24.17
	W	47.12
	O	54.55
	W + O	**56.41**
ESTER	W	**57.35**
	O	55.54
	W + O	55.87

Table 5: Summary of out-of-domain evaluations on two corpora: Rhapsodie (conversatios) and ESTER V2 (broadcast news speech). Best configuration for each corpus is in bold. W is the model from (Muller et al., 2019). Baseline-2 is the best system from the ConLL 2018 shared task, reported from the website.

8 Conclusion

We have detailed the design and evaluation of a model of discourse segmentation for spontaneous, multiparty oral conversation. While segmentation of such data is crucial for recovering discourse structure and for downstream tasks such as meeting summarization, appropriate data sets are lacking and/or diverse. The novelty of our approach lies in transferring a model designed for prepared, written text to conversation transcripts using minimal data or manual annotation. Heuristics or *labeling functions*, which may draw on heterogeneous information sources, are designed by experts and used to *automatically label* training data for supervision.

Our evaluations show that our heuristic-driven approach significantly outperforms a state-of-the-art discourse segmenter trained on prepared, written text. Future work will focus on extending our model to other languages and corpora with varying levels of spontaneity (and disfluency). Weak supervision for NLP is an active subject, and recent new frameworks such as (Lison et al., 2020) might prove even more suitable for sequence tagging problems in future work.

Acknowledgments

This work was supported by the project LinTO, funded by BPI France (No P169201) as part of the *Program d'Investissements d'Avenir*.

It was also partially supported by the ANR (ANR-19-PI3A-0004) through the AI Interdisciplinary Institute, ANITI, as a part of France's "Investing for the Future — PIA3" program, and through the project SUMM-RE (ANR-20-CE23-0017).

Thomas Pellegrini's work is further supported by the LUDAU project (ANR-18-CE23-0005-01).

References

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proc. ICASSP*, volume 1, pages I–1061. IEEE.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Sonia Badene, Kate Thompson, Jean-Pierre Lorré, and Nicholas Asher. 2019. Weak supervision for learning discourse structure. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2296–2305, Hong Kong, China. Association for Computational Linguistics.

Fernando Batista, Helena Moniz, Isabel Trancoso, and Nuno Mamede. 2012. Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. *IEEE transactions on audio, speech, and language processing*, 20(2):474–485.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain.

Christophe Cerisara, Odile Mella, and Dominique Fohr. 2009. JTrans, an open-source software for semi-automatic text-to-speech alignment. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association - Interspeech 2009*, Brighton, United Kingdom.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36, Seattle, Washington. Association for Computational Linguistics.

Viet-Trung Dang, Tianyu Zhao, Sei Ueno, Hirofumi Inaguma, and Tatsuya Kawahara. 2020. End-to-end speech-to-dialog-act recognition. In *Interspeech 2020, 21st Annual Conference of the International*

Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020, pages 3910–3914. ISCA.

Seeger Fisher and Brian Roark. 2007. The utility of parse-derived features for automatic discourse segmentation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 488–495, Prague, Czech Republic. Association for Computational Linguistics.

Christian Fügen, Alex Waibel, and Muntsin Kolss. 2007. Simultaneous translation of lectures and speeches. *Machine Translation*, 21(4):209–252.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, Sapporo, Japan. Association for Computational Linguistics.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. 2009. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, pages 2583–2586. ISCA.

Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue Discourse*, 1(3):1–33.

Javier Iranzo-Sánchez, Adrià Giménez Pastor, Joan Albert Silvestre-Cerdà, Pau Baquero-Arnal, Jorge Civera Saiz, and Alfons Juan. 2020. Direct segmentation models for streaming speech translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2599–2611, Online. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Zhenwen Li, Wenhao Wu, and Sujian Li. 2020. Composing elementary discourse units in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Online. Association for Computational Linguistics.

Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb. 2020. Named entity recognition without labelled data: A weak supervision approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1518–1533, Online. Association for Computational Linguistics.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 147–156, The University of Tokyo, September 24-25, 2010. Association for Computational Linguistics.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text segmentation by cross segment attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4707–4716, Online. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A framework for the analysis of texts. *International Pragmatics Association Papers in Pragmatics*, 1:79–105.

Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts: a surface-based approach. *Computational Linguistics*, 26(3):395–448.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124, Minneapolis, MN. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *The Sixth International Conference on Language Resources and Evaluation*, pages 2961 – 2968, Marrakech, Morocco. ELRA.

Silvia Quarteroni, Alexei V. Ivanov, and Giuseppe Riccardi. 2011. Simultaneous dialog act segmentation and classification from human-human spoken conversations. In *Proc. ICASSP*, pages 5596–5599.

Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason A. Fries, Sen Wu, and Christopher Ré. 2020. Snorkel: rapid training data creation with weak supervision. *VLDB J.*, 29(2-3):709–730.

Alexander J. Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3567–3575.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 228–235.

Caroline Sporleder and Mirella Lapata. 2005. Discourse chunking and its application to sentence compression. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 257–264, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Milan Tofiloski, Julian Brooke, and Maite Taboada. 2009. A syntactic and lexical-based discourse segmenter. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 77–80, Suntec, Singapore. Association for Computational Linguistics.

Xiaolin Wang, Masao Utiyama, and Eiichiro Sumita. 2019. Online sentence segmentation for simultaneous interpretation using multi-shifted recurrent neural network. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 1–11, Dublin, Ireland. European Association for Machine Translation.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Amir Zeldes, Debopam Das, Erick Galani Maziero, Juliano Antonio, and Mikel Iruskieta. 2019. The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 97–104, Minneapolis, MN. Association for Computational Linguistics.

Ruiqing Zhang and Chuanqiang Zhang. 2020. Dynamic sentence boundary detection for simultaneous translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 1–9, Seattle, Washington. Association for Computational Linguistics.

Tianyu Zhao and Tatsuya Kawahara. 2018. A unified neural architecture for joint dialog act segmentation and recognition in spoken dialog system. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–208.

A Descriptions of Labeling Functions

LF name	Polarity	Description	Coverage	Acc.
tony_realturns	1	Predicts a segment boundary (BOS) if ToNy$_W$ predicts a segment boundary at the current token	0.112	0.749
period_bef	1	BOS if the audio+trigram punctuation model predicts a period before the current token, with probability>0.3 and the token is not a filler	0.065	0.712
first_marker	1	BOS if the current token is among a lexicon of specific markers, and is the first of a sequence of such markers (19 markers: *donc* (so), *bon* (well/ok), *ensuite* (next), etc)	0.051	0.677
beg_real_turn	1	BOS if the current token is the first token in a speaker turn (as given to the system)	0.031	0.987
audio_combinations	1	This LF is a set of conditions involving significant audio signal transitions around the current token (changes in pitch or energy), combined with morpho-syntactic constraints on the token and its neighbors (part of speech (pos), syntactic dependencies, pos of the syntactic head)	0.030	0.692
posdep_ngram	1	This LF a set of conditions involving significant morpho-syntactic patterns (pos ngrams, dependency types of the token and its head)	0.025	0.737
keywords	1	BOS if the current token is among a set of opening and acknowledgment markers typical of BOSs that correspond to speaker changes (e.g., *thanks*, *ok*, *good morning*, etc)	0.015	0.591
cconj	1	BOS if the current token figures in a pattern from a specified list of conjunction patterns	0.008	0.887
no_type	0	Predicts no boundary (NO) if the current token falls into a part-of-speech category not typically associated with BOSs	0.331	0.962
no_disfluency	0	NO if the current token is an instance of, or is surrounded by, certain types of disfluencies (e.g., hesitations, repetitions)	0.236	0.923
no_after_markers	0	NO if the current token appears after a series of markers (see first_marker) to avoid segments that consists only of discourse markers	0.075	0.941

Table 6: Description of LFs used by the label model. A few other LFs were tried and dismissed based on analysis of their accuracy on the dev set, and/or the predictions of the generative labeling model on the dev set.