End-to-End Dependency Parsing of Spoken French

Adrien Pupier, Maximin Coavoux, Benjamin Lecouteux, Jérôme Goulian

End-to-End Dependency Parsing of Spoken French

Adrien Pupier, Maximin Coavoux, Benjamin Lecouteux, Jérôme Goulian

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

Motivations

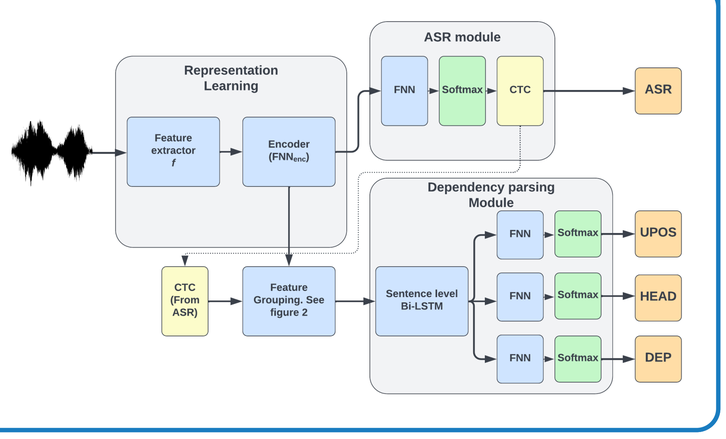
- Is it possible to do syntactic parsing with only the audio modality ?
- Using end-to-end to reduce propagation error.
- Use information (prosodic cues) present in speech to help parsing [4].

Architecture wav2tree

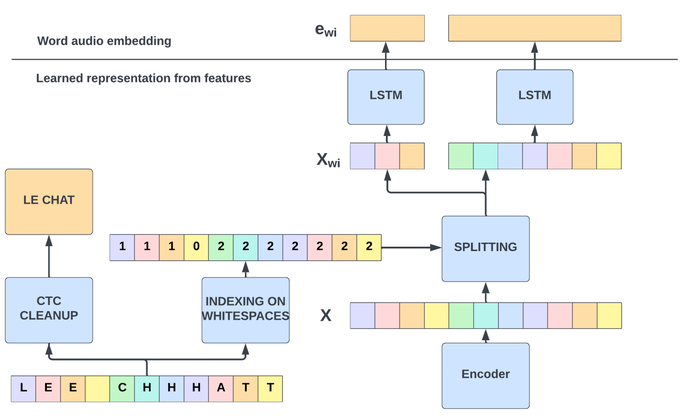
3 modules : Representation learning, ASR and syntactic parsing.

- Feature extractor : MFCC, wav2vec[1], etc. . .
- Compute “word audio embedding" by using the CTC whitespace prediction
- Can use any generic parsing method in the dependency parsing module. We use Dep2Label[5] which reduce the parsing problem to 3 sequence labeling ones. (UPOS, HEAD, DEP)

Global wav2tree architecture



Transforming signal representation into word audio embedding



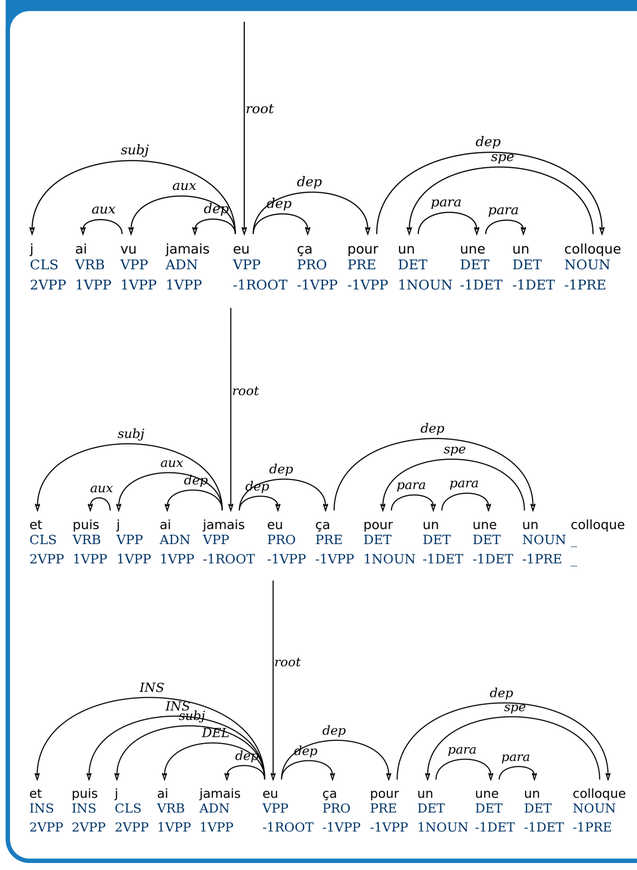
Training and Oracle

- Dependency parsing relies on the segmentation of the signal from the ASR module.
- What to do if the segmentation is wrong during training?

Creation of an oracle :

- Use the gold data and the noisy ASR segmentation to compute the best possible tree
- Allow the model to train on adapted coherent dependency tree
- Heuristics to deal with missing and added token
- Relies on alignment tool

Illustation of the oracle process



Orféo and experimental protocol

- Evaluation of the model on CEFC-Orféo treebank [2].
- Composed of multiple subcorpora with different speech characteristics.

Pipeline baseline :

- Baseline is composed of one ASR module and one dep2Label module.
- The dep2Label module uses `Flaubert-large` embeddings in the pipeline.
- Experiments with and without oracle.

Conclusion

- Using only the audio modality for the syntactic parsing is feasible
- Our model outperforms traditional baseline approaches without using pre-trained language model embedding such as Flaubert.
- Our architecture is heavily customisable and can be used for any task needing word level embedding such as NER, parsing, MWE recognition. . .

future work:

- multimodality text embedding + audio embedding
- Adding language model to improve WER and thus, signal representation segmentation

Composition of Orféo

Corpus	Type	Train size	Test size
Cfpb	spontaneous	3030 (2.2h)	362 (0.3h)
Cfpp	spontaneous	25500 (19.1h)	3232 (2.4h)
Clapi	spontaneous	7682 (5.3h)	967 (0.7h)
Coralrom	spontaneous	10889 (9.6h)	1376 (1.2h)
Crfp	spontaneous	17357 (15.3h)	2259 (2.0h)
Fleuron	spontaneous	1779 (1.4h)	217 (0.2h)
Oral-Narrative	prepared/read	8388 (7.3h)	1050 (1h)
Ofrom	spontaneous	11665 (9.3h)	1476 (1.2h)
Reunions	spontaneous	10067 (8.0h)	1245 (1h)
Tcof	spontaneous	16063 (11.6h)	1997 (1.5h)
Tufs	spontaneous	35990 (24.3h)	4525 (3.0h)
Valibel	mixed	21095 (17.5h)	2753 (2.3h)
Total	mixed	169505 (130.9h)	21459 (16.8h)

Results

Corpus	Wav2Tree				Pipeline				Pipeline + oracle			
Input	SIGNAL				Train : Gold - Test : ASR				Train : ASR - Test : ASR			
Pre-trained	Wav2vec2[3]				Wav2vec2 + flauBERT				Wav2vec2 + flauBERT			
Parameters	350M+33M				350M+373M+32M				350M+373M+32M			
	WER	UPOS	UAS	LAS	WER	UPOS	UAS	LAS	WER	UPOS	UAS	LAS
Cfpb	36.7	77.6	**73.4**	**68.8**	**35.0**	76.6	71.8	67.5	**35.0**	**78.0**	73.0	68.5
Cfpp	40.6	**71.8**	**66.6**	**61.8**	**40.4**	69.9	64.8	60.2	**40.4**	70.7	65.0	60.4
Clapi	62.3	**58.8**	**54.4**	**47.6**	**62.0**	56.4	53.5	46.7	**61.9**	57.4	53.75	47.0
Coralrom	26.0	**83.9**	**77.8**	**74.5**	**25.4**	82.0	75.3	71.8	**25.4**	83.0	75.9	72.4
Crfp	30.9	**81.7**	**75.9**	**72.2**	**30.0**	80.3	74.0	70.4	**30.0**	81.3	74.8	71.1
Fleuron	39.9	71.9	65.3	60.5	**38.9**	71.0	65.1	**61.3**	**38.9**	**72.1**	**66.0**	61.1
Oral-Narrative	12.3	**93.2**	**87.9**	**85.7**	**11.4**	92.1	86.1	83.4	**11.4**	93.0	86.6	84.2
Ofrom	22.1	**85.2**	**79.3**	**75.9**	**21.3**	84.2	77.9	74.7	**21.3**	85.1	78.4	75.1
Reunions	46.9	**67.7**	**61.8**	**56.3**	**45.8**	65.5	60.3	55.1	**45.8**	66.7	60.8	55.6
Tcof	38.9	**74.3**	**67.4**	**62.7**	**38.3**	72.3	65.4	60.8	**38.2**	73.2	65.6	61.0
Tufs	37.6	**75.2**	**69.6**	**65.1**	**36.9**	73.5	67.8	63.5	**36.9**	74.6	68.7	64.1
Valibel	26.1	**82.8**	**76.9**	**73.2**	**25.4**	81.3	75.4	71.7	**25.4**	82.2	75.7	71.8
Orféo full	35.0	**77.4**	**71.7**	**67.5**	**34.0**	75.8	70.0	65.8	**34.0**	76.7	70.5	66.2

WER: Word Error Rate • **UPOS**: Universal Part of Speech • **UAS**: Unlabeled attachment score • **LAS**: labeled attachment score

References

[1] Alexei Baevski et al. “wav2vec 2.0: A framework for self-supervised learning of speech representations”.

[2] Christophe Benzitoun et al. “Le projet ORFÉO: un corpus d’étude pour le français contemporain”.

[3] Solène Evain et al. “LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech”.

[4] Patti J Price et al. “The use of prosody in syntactic disambiguation”.

[5] Michalina Strzyz et al. “Viable Dependency Parsing as Sequence Labeling”.

Acknowledgement

The author acknowledges the support of the French Agence Nationale de la Recherche (ANR) PROPICTO under reference ANR-20-CE93-0005

Contact: adrien.pupier@univ-grenoble-alpes.fr