Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2605–2614
Marseille, 11–16 May 2020

Evaluating Sentence Segmentation in Different Datasets of Neuropsychological Language Tests in Brazilian Portuguese

Edresson Casanova[1], Marcos V. Treviso[2*], Lilian C. Hübner[3], Sandra M. Aluísio[1]
[1]University of São Paulo, [2]Instituto de Telecomunicações, [3]Pontifical Catholic University of Rio Grande do Sul
edresson@usp.br, marcos.treviso@lx.it.pt, lilian.hubner@pucrs.br, sandra@icmc.usp.br

Abstract

Automatic analysis of connected speech by natural language processing techniques is a promising direction for diagnosing cognitive impairments. However, some difficulties still remain: the time required for manual narrative transcription and the decision on how transcripts should be divided into sentences for successful application of parsers used in metrics, such as Idea Density, to analyze the transcripts. The main goal of this paper was to develop a generic segmentation system for narratives of neuropsychological language tests. We explored the performance of our previous single-dataset-trained sentence segmentation architecture in a richer scenario involving three new datasets used to diagnose cognitive impairments, comprising different stories and two types of stimulus presentation for eliciting narratives — visual and oral — via illustrated story-book and sequence of scenes, and by retelling. Also, we proposed and evaluated three modifications to our previous RCNN architecture: (i) the inclusion of a Linear Chain CRF; (ii) the inclusion of a self-attention mechanism; and (iii) the replacement of the LSTM recurrent layer by a Quasi-Recurrent Neural Network layer. Our study allowed us to develop two new models for segmenting impaired speech transcriptions, along with an ideal combination of datasets and specific groups of narratives to be used as the training set.

Keywords: Sentence Segmentation, Impaired Speech, Neuropsychological Language Tests

1. Introduction

Language assessment has been shown to be an efficient complementary tool for detecting cognitive and neuropsychological disorders, therefore present in most tests, tasks and batteries that evaluate cognitive processes. For example, neuropsychological language tests are an important tool for diagnosing individuals with significant depression in Alzheimer's disease (AD) (Fraser et al., 2016), to differentiate between Mild Cognitive Impairment (MCI) and AD (Drummond et al., 2015), to differentiate between AD and other neurodegenerative dementias (Yancheva et al., 2015; Beltrami et al., 2018) and to differentiate variants of neurodegenerative dementias, such as in Primary Progressive Aphasia (PPA) (Fraser et al., 2014).

Language assessment has been performed mainly by using discursive production in which narratives are largely used, since they are a natural form of communication and favor the observation of the patient's functionality in everyday life (Tillas, 2015). The discourse tasks used to assess the narrative productions of elder individuals are often based on: (i) an illustrated story book without a text (e.g. Cinderella), (ii) an immediate and delayed retelling of a story orally presented, or (iii) a single scene or a sequence of scenes, presented on pictures, of a common event that occurs in daily life.

With regard to specific batteries used to evaluate language in discourse tasks, we can cite a few (Wechsler, 1997; Bayles and Tomoeda, 1993; Goodglass et al., 1983; Hübner et al., 2019). Discourse tasks that require some degree of memorization are usually included in verbal memory tests. This is the case of the Logical Memory Subtest task from the Wechsler Memory Scale, used for assessing episodic memory (Wechsler, 1997). In this task, an individual reproduces a story immediately after listening to it (immediate recall); thirty minutes later, subjects are asked to recall the story again (delayed recall). The retellings are transcribed for further analysis. The higher the number of recalled elements of the narrative, the higher the memory score. This procedure is also used in the Arizona Battery for Communication Disorders of Dementia (ABCD) (Bayles and Tomoeda, 1993). The single-scene description task called "The Cookie Theft Picture" is part of the Boston Diagnostic Aphasia Examination (BDAE) (Goodglass et al., 1983). The Cookie-Theft picture has been shown to be clinically relevant in identifying linguistic deficits in Alzheimer's disease patients given the importance of using visual stimuli when evaluating individuals of this group. The Cinderella story is also very widely used in the assessment of aphasia and some types of dementia. (Fraser et al., 2015) and (Aluisio et al., 2016) based their work on the story of Cinderella. Participants were given a sequenced picture book (without words) to remind them of the story; then they were asked to tell the story in their own words. The narrative samples were then transcribed by trained annotators.

A challenge in choosing the type of neuropsychological assessment of individuals with AD and MCI is the use of a battery that can distinguish these individuals. ABCD appears as an option since it is capable of detecting mild stage AD, while the *Bateria de Avaliação da Linguagem no Envelhecimento* (BALE) (Battery of Language Assessment in Aging, in English) (Hübner et al., 2019) was developed for application to patients with different educational levels, including illiterate ones, a very common condition among Brazilian elders.

All the studies cited above confirm that the automatic analysis of connected speech by natural language processing techniques (NLP) is a promising direction for diagnosing cognitive impairments. However, some difficulties still re-

*Work carried out during the master's course at the University of São Paulo.

main: the time required for manual narrative transcription and the decision on how transcripts should be divided into sentences for (i) extracting narrative recall scores automatically from semantic similarity methods applied to sentences — the shorter the sentences, the better the method response (see (Borges dos Santos and Aluísio, 2020)) and (ii) the successful application of parsers used in metrics to analyze the transcripts. One of these metrics is called Idea Density and was originally proposed as a way of measuring the memory load of narratives, by representing the underlying content of the text as a series of semantic units, called propositions or ideas. The method proposed by da Cunha et al. (2015) is a rule-based system acting upon dependency trees, strongly depending on a robust parser.

Growing consensus in the NLP area indicates that in order to have fully automated systems for diagnosing cognitive impairments, an NLP pipeline must use an Automatic Speech Recognition (ASR) system. Although to build a high performance ASR for pathological language can be a long term research, we will ignore the issue (i) for now. If we have a manual transcription or an automatically generated one, we still have to detect sentence boundaries, therefore, we will focus on this task in this paper.

Since the majority of studies on diagnosing cognitive impairments by NLP methods deal with English-speaking patients (Filiou et al., 2019), in this study we will evaluate the Brazilian Portuguese (BP) language in order to contribute with datasets and studies to develop automatic analysis of connected speech in BP. Our motivation for this study was to explore the performance of a single-dataset-trained sentence segmentation architecture in a richer scenario involving three new datasets. Therefore, here, we evaluate four datasets used to diagnose cognitive impairments (see Section 3), comprising different stories and two type of stimulus presentation for eliciting narratives: (i) oral stimuli presentation with retelling, where sequencing discourse marks, such as “e”,“aí”, “daí” and “então” (and, then, in English) and confirmatory discourse marks “né” and “ok” (ok, in English) are frequent and (ii) visual stimuli, via both illustrated story book and sequence of scenes of a common event, where deictic expressions (place deixis) are pervasive, such as “aqui” and “aí” (here) and “ali” and “lá” (there), besides presenting sequencing discourse marks and confirmatory discourse.

Figure 1 shows the result of a manual transcription in BP of a narrative of the ABCD story telling task, which presents a story about a woman who is unaware of having lost her wallet while doing the shopping; she then receives a call from a little girl who found the wallet. As we can see in (a) the transcript without punctuation prevents the direct application of NLP methods that rely on sentence segmentation for the correct use of tools as taggers and parsers. These tools are used to implement metrics of syntactic complexity, basic counts of PoS tags and to analyze other levels of language to diagnose cognitive impairments.

When the architecture developed in the project DeepBonDD[1] was trained with The Cinderella Story dataset (a production task elicited via an illustrated story book) and evaluated with the other three datasets analyzed in this paper, F_1 values were much lower than the original work on Cinderella (Treviso et al., 2017a) (Table 1). The average of F_1 in the three datasets, for all classes, is 0.59. In the original evaluation with training and cross-validation testing in the same dataset, the best F_1 value for Controls was 0.76, for MCIs, 0.74, and for ADs, 0.66. However, Table 1 shows that, in general, for sentence segmentation, more data is beneficial, independently of task and topic of datasets.

(a) **ahm** uma senhora **foi fazer compras no me** foi no mercado não lembrava o local **no me** fazer compras e quando ela foi pagar a conta no caixa percebeu que estava sem a carteira aí ela **foi deixou a mercadoria** não levou a mercadoria voltou para casa chegando em casa toca o telefone era uma garotinha avisando ela que que tinha achado a carteira **é isso tem mais coisa não cortei eu resumi o que eu ouvi**

(b) ahm uma senhora foi fazer compras no me foi no mercado. não lembrava o local . no me fazer compras . e quando ela foi pagar a conta no caixa percebeu que estava sem a carteira . aí ela foi deixou a mercadoria . não levou a mercadoria . voltou para casa . chegando em casa toca o telefone . era uma garotinha avisando ela que que tinha achado a carteira . é isso . tem mais coisa . não cortei . eu resumi o que eu ouvi .

Figure 1: (a) Narrative transcribed where there is no punctuation or capitalization, besides presenting several disfluencies, such as unlexicalized filled pauses, restarts and patient’s comments, shown in bold. (b) Narrative manually segmented of a retelling task using The Wallet Story.

Given this motivation scenario, where the main goal was to develop a robust and generic segmentation system for narratives of neuropsychological language tests, the present study tries to answer three questions:

1. Would modifications to the Recursive Convolutional Neural Networks (RCNN) architecture proposed by (Treviso et al., 2017a) have better performance on sentence boundary detection of new tasks and story topics? (Section 2 presents details about the RCNN architecture proposed by (Treviso et al., 2017a) and Section 4 presents the three modifications evaluated here);

2. Are there particularities in tasks that elicit narratives with visual stimuli not present on those elicited with oral stimuli (and vice versa), requiring specific sentence segmentation detectors for each task? (Section 5 presents our experiments on this issue.)

3. Do the story topics of language tests and the group of elders of a dataset negatively impact sentence segmentation detectors, also requiring specific sentence segmentation detectors? (Section 6 presents our evaluations on this issue.)

Section 2 presents a literature review on sentence segmentation, focusing on spontaneous and impaired speech. Section 7 presents the discussions and the contributions towards building a full-fledged system for automating neuropsychological tests in Portuguese.

2. Related Work on Sentence Segmentation for Impaired Speech

Due to the increasing usage of ASR systems, which usually output a stream of tokens without any capitalization

[1] https://github.com/mtreviso/deepbond

Training set	Test set	MCIs	Controls	ADs	Average
Cinderella - Same class	The Dog Story	0.43	0.54	0.56	0.51
Cinderella - All classes	The Dog Story	**0.58**	0.59	0.54	0.57
Cinderella - Same class	Lucia	n/a	0.67	0.54	0.60
Cinderella - All classes	Lucia	n/a	**0.66**	**0.62**	0.64
Cinderella - Same class	Wallet	0.54	0.56	n/a	0.55
Cinderella - All classes	Wallet	0.57	0.53	n/a	0.55

Table 1: Robustness tests in terms of F_1 using the original RCNN model trained on the Cinderella dataset. "Same class" means that the method was trained only with the same specific class used for testing. "All classes" means that data from all classes were used for training. "n/a" entries denote that the tested dataset does not have examples for that specific class.

or punctuation symbols, methods for detecting sentences boundaries were applied to solve the need of subsequent tools (such as taggers and parsers) in an NLP pipeline.

Previous methods, such as Decision Trees combined with Language Models (Shriberg et al., 2000; Liu et al., 2005b; Christensen et al., 2006), Maximum entropy models (Batista et al., 2012) and CRFs (Khomitsevich et al., 2015; Fraser et al., 2015) were applied both for prepared and spontaneous speech. These methods rely on lexical and prosodic clues (e.g. pitch, energy and pauses) in order to detect the correct position of a sentence boundary. For instance, the CRF method proposed by (Fraser et al., 2015) uses lexical, prosodic and Part-of-Speech (PoS) tags as features to segment speech from elder people with aphasia. They found that by using all these features together the model yields better results and the mistakes made by the model don't affect much the syntactic structure of the segmented transcript.

More recently, Recurrent and Convolutional Neural Networks were employed for both types of speech and achieved good results by using word embeddings as the lexical representation of words (Tilk and Alumäe, 2016; Che et al., 2016; Treviso et al., 2017a; González-Gallardo and Torres-Moreno, 2018), suggesting that deep neural networks can be successfully applied for this task.

Prosodic features have been shown to be very effective to discriminate between different types of sentence boundaries and in general their usage reflects better results (Shriberg et al., 2009; Huang et al., 2014; Khomitsevich et al., 2015). However, to put prosodic features into practice we need alignments between the audio and its transcription, which is hard to obtain mainly due to the low quality of the recordings. This problem is even more critical for impaired speech, where patients with cognitive impairment usually produce a narrative in which sentences are not syntactically well-formed, words are pronounced in a way that modifies their original morphology, and utterances have low prosody quality (elder speakers with a very low voice volume). Even long pauses are not always an indication of sentence boundaries due to word-finding difficulty of elders (Fraser et al., 2015). Therefore, prosody is hardly ever a good feature for the classifier.

Recent studies show that by using only lexical clues it is possible to achieve a comparable performance with methods that use prosodic features altogether (Klejch et al., 2016; Klejch et al., 2017). Moreover, by leveraging transfer learning techniques it is possible to reduce the drop in performance even more. For example, the method developed by (Treviso et al., 2017a), which was evaluated with different types of word embeddings, showed that by using a good word embedding representation it is possible to achieve similar results in the SOTA. Their method consists of a combination of Recurrent and Convolutional Neural Networks (RCNN). Its complete architecture is composed by the following four components:

1. An embedding layer maps words to dense vectors representations;
2. These vectors are fed to a convolutional layer that is responsible for the automatic extraction of new features depending on neighboring words;
3. The new extracted features are passed to a Bi-LSTM to capture long range dependencies; and
4. The output of the recurrent layer is projected to a binary output where a softmax operation is calculated, giving the probability of whether or not the word precedes a sentence boundary.

Since the number of sentence boundaries is much lower than the one of non-sentence boundaries, this is categorized as an unbalanced classification problem. To deal with this, the RCNN gives a higher weight to the minority class in the objective function.

The main drawback of the RCNN model is its large number of parameters combined with the small amount of training data (usual in clinical data), which usually leads to overfitting, and therefore careful regularization strategies have to be employed. In practice, we found that for narratives of new story topics the RCNN model is not able to often detect good sentence boundaries, relying on discourse marks and therefore generating very small sentences, with a main verb as the manual segmentation does.

By inspecting the errors of the RCNN model, we also found that its most common mistakes are related to deictics, sequencing and confirmatory marks preceding (e.g. "lá", "né", "ok") and succeeding sentence boundaries (e.g. "aí", "daí", "então"). Although it was shown that these errors do not affect too much the syntactic structure of the sentence, they could be easily captured by considering lexical clues more effectively (Treviso et al., 2018). Examples of places where the model should have put a sentence boundary but it missed it (i.e. false negatives) are shown in Fig. 2

(a) menino que foi na cidade . aí tá caminhando na rua . daí viu as pessoas lá . daí encontrou um cachorrinho . o cachorrinho tava perdido . chegando lá a mãe abriu a porta . e ele pediu pra mãe deixar o cachorro lá . morar com ele lá . aí arrumou até uma casinha pro cachorrinho . aí ela consentiu ele deixar até fazer uma casinha pro cachorro .

(b) menino que foi na **cidade aí** tá caminhando na **rua daí** viu as pessoas **lá daí** encontrou um **cachorrinho o** cachorrinho tava perdido . chegando lá a mãe abriu a porta . e ele pediu pra mãe deixar o cachorro **lá morar** com ele lá . aí arrumou até uma casinha pro cachorrinho . aí ela consentiu ele deixar até fazer uma casinha pro cachorro .

Figure 2: (a) Manual segmentation of a narrative elicited via a sequence of scenes of a common event (The Dog Story). (b) Example of errors made by the RCNN model; slots where a sentence boundary should have been put are shown in bold.

3. Datasets

Four datasets were used to train our models (Sections 3.1, 3.2 and 3.3). As a preprocessing step we removed capitalization information and in order to simulate high-quality ASR, we left all speech disfluencies intact. Demographic information of participants and statistics about the narratives of our study are presented in Table 2. Datasets of Neuropsychological Language Tests are typically small, as can be seen in Table 2. Table 2 shows the uniform mean length of sentences of three datasets (The Wallet Story, The Dog Story and The Cinderella Story), with regards to the groups MCI and Controls. This is an interesting feature to train/test a model, using a large dataset which combines several stories. Cinderella's mean length of narratives is very long, while both retellings produce short narratives.

In the four datasets of this study, we have segmented sentences using prosodic, syntactic and semantic knowledge, to create short sentences, with a sole idea, i.e. with an unique main verb. Therefore, coordinated sentences where divided. Although this decision has an impact on certain syntactic metrics, such as the number of sentences with coordination and the length of sentences, it makes possible for parsers to function properly over impaired speech. The manual sentence segmentation was performed by peers in two datasets, and both kappa values are very high (Landis and Koch, 1977): the kappa value for the Cinderella Story was 0.84 (almost perfect agreement), and for the Dog Story was 0.77 (substantial agreement). Therefore, the remaining annotation was performed by a sole annotator.

3.1. The Wallet Story from ABCD

ABCD is a standardized test battery for the comprehensive assessment and screening of dementia. It includes 17 subtests that evaluate linguistic expression, linguistic comprehension, verbal episodic memory via immediate/delayed recall of stories, visuospatial construction, and mental status. The subtest which is important for our study is the evaluation of the episodic memory, which is composed of the immediate and late retelling of a memorized story from (Bayles and Tomoeda, 1993), the Wallet Story. This story was translated and adapted to BP by Danielle Rüegg, Isabel Maranhão de Carvalho, Leticia Lessa Mansur and Márcia Radanovic, and was administered and collected by the team coordinated by Professor Dr. Leticia Lessa Mansur at the University of São Paulo Medical School to 23 elders with MCI and 12 healthy aging adults; totaling 70 narratives. This test has 17 units of information, with possible alternatives, with 17 being its maximum score.

3.2. The Cinderella Dataset

The Cinderella dataset consists of spontaneous speech narratives produced during a test to elicit narrative discourse with visual stimuli, using a book composed of sequenced pictures portraying the the Cinderella Story. In the test, the examinee verbally tells the story to the examiner based on the pictures. The narrative is recorded and manually transcribed by a trained annotator who scores the narrative by counting the number of recalled propositions/units of information; there are 28 informational units to be recalled, presented in 23 pictures. This dataset consists of 60 narratives from BP speakers (20 controls, 20 with AD, and 20 with amnestic MCI), diagnosed and collected at the University of São Paulo Medical School and also used in (Toledo et al., 2017; Aluisio et al., 2016; Treviso et al., 2017b; Treviso et al., 2017a).

3.3. The Dog Story and Lucia Story Datasets from BALE

BALE is a standardized battery with norms for the healthy elders Brazilian population illiterate, with low (2 to 8 years of schooling) and high (9 years or more) education, from 60 to 90 years old, described in (Hübner et al., 2019). BALE provides the academy and clinicians with standardized and validated tasks, filling an important gap in terms of tasks validated for BP, specially at the discourse level. It was conceived by the adaptation of other tasks nationally and internationally used to test language impairment mainly in AD, following psycholinguistic criteria, including imageability, frequency, animability, extension, among others, such as cultural issues. It consists of 10 linguistic tasks, assessing from the word level, in the naming task, for example, to the discourse level. One of its differentials is to evaluate discourse in four types of narrative texts, especially at the production level, but with the implicit textual comprehension as well. This battery was chosen because its aim is to allow for its administration to elder people who are illiterate and/or of low educational level, who represent the majority of the aged sample assisted by the public health system in Brazil. The Dog Story and Lucia Story are two of the four narrative texts from the BALE instrument. The Dog Story dataset is composed of transcriptions from the oral narrative production test based on the presentation of a set of seven pictures telling a story of a boy who hides a dog that he found on the street, based on the story of LeBooeuf (Le Boeuf, 1976). This dataset consists of 106 narrative texts from BP speakers, including 82 healthy aging adults, 12 with AD, and 12 with MCI. BALE also includes a task of retelling and text comprehension of an orally presented story called Lucia Story. This test has 24 units of information, with possible alternatives, with 24 being its maximum score. This retelling test was administered to 9 Alzheimer's individuals and 80 healthy aging adults. Both datasets were collected by the team coordinated by Professor Dr. Lilian Cristine Hübner of the School of Humanities at the Pontifical Catholic University of Rio Grande do Sul (PUCRS).

Stories	Groups	Nb. of Subjects	Age	Years of Education	Nb. of Sentences	Mean length of Sentences (σ)	Mean length of Narratives (σ)
The Wallet Story	MCI	23	62+	4+	376	7.45 ($\pm$3.99)	60.87 ($\pm$17.22)
	Control	12	55+	4+	184	7.70 ($\pm$4.29)	59.00 ($\pm$14.41)
The Lucia Story	Control	80	63+	2+	564	8.44 ($\pm$5.57)	59.51 ($\pm$21.36)
	AD	9	68+	1+	39	6.54 ($\pm$5.59)	28.33 ($\pm$18.37)
The Dog Story	MCI	12	57+	2+	173	8.26 ($\pm$4.43)	119.08 ($\pm$41.61)
	Control	82	60+	0+	1170	8.44 ($\pm$5.10)	120.46 ($\pm$51.65)
	AD	12	59+	0+	153	7.60 ($\pm$5.50)	96.92 ($\pm$37.56)
The Cinderella Story	MCI	20	60+	3+	618	12.38 ($\pm$7.40)	404.80 ($\pm$198.40)
	Control	20	60+	3+	654	12.79 ($\pm$7.23)	395.25 ($\pm$210.33)
	AD	20	60+	3+	794	9.83 ($\pm$7.00)	390.30 ($\pm$285.91)

Table 2: Statistics of narratives and of the Control and patient groups; the first two datasets are based on retellings and the last ones are based on sequenced figures.

4. Exp. I: Would New Architectures Have Better Performance?

Here, we propose and evaluate three modifications to the RCNN architecture developed by (Treviso et al., 2017a), namely: (i) the inclusion of a Linear Chain CRF to capture pairwise dependencies between labels; (ii) the inclusion of a self attention mechanism with the aim of capturing very long dependencies; and (iii) the replacement of the LSTM recurrent layer by a Quasi-Recurrent Neural Networks (QRNN) (Bradbury et al., 2016) layer in order to reduce the number of trainable parameters.

A CRF model can be helpful since we have sequences of labels that are very unlikely (or even impossible) to happen, such as a sequence of three sentence boundaries one after the other: `B B B`. Furthermore, by applying Viterbi decoding we can seek the best sequence of labels taking into account these transition likelihoods. In contrast to RNNs that have to remember decisions across a very long stream of tokens, attention mechanisms can access distant positions in the input at any moment, therefore they can be very helpful to learn very long dependencies. Despite having less parameters than LSTMs, a QRNN layer is based on convolutional and pooling operations, which can be computed effectively in parallel and, as a result, decrease both training and inference time.

To choose the best architectures, several configurations were evaluated using greedy search on the hyperparameters found in Table 3. We used a 5-fold cross-validation on the MCI class set of The Cinderella Story dataset. The choice of evaluating only on the MCI class was due to the high demanding time of running all experiments in all four datasets and the fact that this class represents cognitive impairment between the characteristics of Controls and ADs narratives. As we can see in Table 3, the models explore the use of CNNs, RNNs, QRNNs, different variants of attention mechanisms, and CRF; mixed models with two or more combinations were also explored. The dot product attention is the scaled version proposed by (Vaswani et al., 2017); the general attention is also known as Luong attention (Luong et al., 2015); the additive attention is also known as Bahdanau attention (Bahdanau et al., 2015).

Hyperparameters	Values
Conv. filters	35, 50, 100, 200
Kernel size	1, 3, 5, 7
Conv. dropout	0.0, 0.25, 0.5, 0.75
Recurrent hidden size	35, 50, 100, 200
Recurrent type	RNN, GRU, LSTM, QRNN
Recurrent dropout	0.0, 0.5
Attention dropout	0.0, 0.25, 0.5, 0.75
Attention variant	Dot Product, General, Additive
Attention hidden size	35, 50, 100, 200, 300
Number of heads	1, 2, 4
Multi-head hidden size	50, 100, 200

Table 3: Hyperparameters tried during greedy search.

We trained our models for a maximum of 40 epochs using small batch sizes. We found that smaller batch sizes work better in practice for models that have a CRF at the end, therefore we set the batch size to 1 for all configurations in order to have comparable results. We employed early stopping with patience of 5 epochs. We optimized the model's parameters using the Adam optimizer with the weight decay fix implementation (Loshchilov and Hutter, 2019). In order to avoid overfitting, we also used ℓ_2 regularization with $\lambda = 0.01$. For all other optimizers hyperparameters, we used the default ones defined on the PyTorch implementation. Finally, in all our experiments we used pre-trained word embeddings from (Treviso et al., 2017a), selecting the 600D Word2vec-skipgram type (due to its higher performance) and keeping them freezed during training.

Taking into account all configurations, more than 200 new architectures were trained and evaluated on the MCI class set of the Cinderella dataset. Since in this work we have more training data from different language tests, our main aim was that we can have a new model that performs better than the RCNN and can also generalize better to datasets of different story topics.

Due to space constraints we do not report the results for each configuration. Nevertheless, these experiments showed that dropout after convolutional and recurrent lay-

Model	Retelling			Sequenced Figures			Average
	MCIs	Controls	ADs	MCIs	Controls	ADs	
1. CRF	0.38	0.44	0.36	0.63	0.68	0.76	0.54
2. QRCNN	0.71	0.76	0.64	0.88	0.84	0.85	0.78
3. RCNN (original)	0.72	0.76	0.65	0.88	0.85	**0.91**	0.79
4. RCNN + CRF	0.74	0.76	0.65	0.88	0.85	0.85	0.79
5. CNN	0.72	0.76	0.65	0.88	0.85	0.85	0.79
6. CNN + CRF	0.72	0.76	0.63	0.88	0.85	0.83	0.78
7. CNN + ATTN	0.72	0.75	0.65	0.88	0.83	0.86	0.78
8. CNN + ATTN + CRF	0.74	0.76	0.65	**0.89**	0.84	0.83	0.79
9. RNN	0.73	0.76	0.64	**0.89**	0.85	0.88	0.79
10. RNN + CRF	0.76	0.78	0.66	**0.89**	**0.86**	0.85	0.80
11. RNN + ATTN	0.71	0.74	0.64	0.88	0.85	0.85	0.78
12. RNN + ATTN + CRF	**0.77**	**0.79**	**0.67**	**0.89**	**0.86**	0.85	**0.81**

Table 4: Cross-validation F_1 scores for each method on both retelling and sequenced figures datasets.

ers is an important factor to prevent overfitting. The dropout rate after convolutional layers was usually set to 0.25, and 0.5 for recurrent layers. For all models with convolutional layers, we found that the best kernel size was 7, and the best number of filters varied between 100 and 200. In general, the number of recurrent units varied between 100 a 200, except for models based on QRNNs, which performed better with 50 units. As for the recurrent unit type, we found that LSTMs usually perform better than GRUs and QRNNs. Finally, the general and additive attention variants were the ones that yielded the best results.

In order into train the best architectures with the datasets presented in Table 2, datasets of tasks that elicit narratives with visual stimuli (Cinderella and Dog Story) were joined and used for training the selected architectures. The datasets of tasks that elicit narratives with oral stimuli (Lucia and Wallet Story) were also joined and used for training the same architectures.

Taking in account the previous experiment with the MCI class set of The Cinderella Story dataset, we selected 12 different models (including the original RCNN) with their best configuration. Moreover, in order to show the impact of each architecture, we chose models that have unique configuration of layers. Table 4 shows 10-fold cross-validation results for datasets based on retellings and based on sequenced figures using these selected models.

With the exception of the CRF model, all others yield similar results, ranging from 0.78 to 0.81 F_1 score in average, with the best value being the RNN + ATTN + CRF model. Therefore, our answer to the question "Would new architectures have better performance?" is no, there was no significant increasing in F_1 score. It is a fact that more data taken from similar distribution (same task of a language battery) is beneficial. The RNN + ATTN + CRF model also makes a small improvement on the F_1 score when compared to the original RCNN model; however, the RCNN is still very competitive. Finally, it is worth noticing that all architectures use the same word embeddings extracted from (Treviso et al., 2017a), which were pre-trained using a large collection of written texts and, as a consequence, some of the lexical clues for sentence boundaries of spontaneous speech were probably not captured by this representation.

5. Exp. II: Does the Task Require a Specific Sentence Segmentation Detector?

An important question is whether models trained in a same task also generalize well to the other task being evaluated, i.e., can we use a model trained on retellings to segment narratives based on sequenced figures and vice versa?

To answer these questions, we chose 3 top recurrent models (9, 10 and 12) from the cross-validation experiments of Table 4 with addition of the original RCNN.

To avoid an unfair comparison between the datasets, the models were then trained again via 10-fold cross-validation where we also randomly split the other dataset into 10 folds to be used for testing. In order to evaluate each model, we first train the model using the training set of each fold and evaluated it twice, once using its test set counterpart and the other time using the respective fold of the other dataset. For example, if we trained the model for the 1st fold of sequenced figures, we tested it both on (i) its test subset; (ii) the 1st fold of the retelling dataset. Since we do not have a validation set, we can not employ an early stopping procedure, so instead we estimate a good number of epochs based on the average number of epochs obtained in the experiments reported on the Table 4.

Table 5 presents F_1 scores of our four best models from the previous experiment trained on the retelling datasets, and Table 6 presents F_1 scores of our four best models from the previous experiment trained on the sequenced figures datasets. In both tables the models were tested on both retelling and sequenced figures datasets.

Table 6 shows the best generalization result: the model RNN + ATTN + CRF. It was trained on sequenced-figures datasets and presents the best average values in both testing data (retellings and sequenced figures as well). The model RNN, in Table 6, also presents similar behavior. From these results, we can assume that sequenced-figures narratives bring linguistic features also present in retellings, but the reverse direction is not true, as we can see in Table 5. However, if a researcher will only work on retelling tasks, Table 5 shows that using only retelling datasets for training led to

Model	Retelling				Sequenced Figures				Average
	MCIs	Controls	ADs	Average	MCIs	Controls	ADs	Average	
RCNN (original)	0.84	0.80	0.75	0.80	0.51	0.60	0.44	0.52	0.66
RNN	0.72	0.77	0.61	0.70	**0.66**	**0.74**	**0.60**	**0.67**	**0.68**
RNN + CRF	0.84	**0.83**	**0.81**	**0.83**	0.51	0.57	0.41	0.50	0.66
RNN + ATTN + CRF	**0.85**	**0.83**	0.74	0.81	0.51	0.58	0.41	0.50	0.65

Table 5: F_1 scores of our best models trained on the retelling datasets and tested on both datasets.

Model	Sequenced Figures				Retelling				Average
	MCIs	Controls	ADs	Average	MCIs	Controls	ADs	Average	
RCNN (original)	0.68	0.74	0.61	0.68	0.67	0.70	**0.67**	**0.68**	0.68
RNN	0.69	0.74	**0.64**	0.69	**0.69**	0.71	0.63	**0.68**	0.68
RNN + CRF	**0.73**	**0.76**	0.61	**0.70**	0.67	0.71	0.61	0.66	0.68
RNN + ATTN + CRF	0.70	**0.76**	0.63	**0.70**	**0.69**	**0.72**	0.62	0.67	**0.69**

Table 6: F_1 scores of our best models trained on the sequenced figures datasets and tested on both datasets.

better results for the retelling task.

6. Exp. III: Does the Story Topic Demand a Specific Segmentation Detector?

Here, we evaluate if we can generalize on the topic of stories used in language batteries, allowing the creation of a generic and unique model for the sentence segmentation task for impaired speech transcriptions. Table 7 shows the results of our best models presented in Table 5 and 6 trained here with three datasets and tested with the fourth remaining dataset, totaling 16 models, and allowing a rich combination to evaluate the best results for the segmentation task. We show in bold the three best average values of F_1 scores (models 8, 9 and 10) (Avg 1). We also calculated the average of F_1 by model (RCNN, RNN, RNN+CRF and RNN+ATTN+CRF) and training class (same, all and MCI and Control classes) (Avg. 2), allowing us to find the model with the best generalization for new data, independently of the task.

Model 8 (RNN + ATTN + CRF) uses the datasets Cinderella, Dog and Wallet for training and Lucia for testing. The training was done with two datasets of narratives elicited by visual stimuli (Cinderella and Dog Story), which have already been selected as the best stimuli to generalize the best model in Experiment 2, taking in account the training with sequenced figures datasets (Table 6).

Considering a new dataset, independently of its task, the best model is the RNN, trained with "All groups" (Avg. 2). Comparing its value of F_1 (0.66) with the results of Table 1 (our previous generic segmentation system), there was a increase of 0.7 in the F_1 score.

Taking all these results into consideration, we chose the model RNN + ATTN + CRF for creating detectors for a specific task, i.e. training with datasets of tasks that elicit narratives with visual stimuli (Cinderella and Dog Story) and with datasets of tasks that elicit narratives with oral stimuli (Lucia and Wallet Story). Also, we chose to use only the groups of Controls and MCIs as their narratives are more similar than those of DAs. The RNN + ATTN + CRF model returned the best results in Experiments 1 and 2. Models with CRF generally do not generate two boundaries in sequence, as in "Ela saiu de. casa.", since the constraint of two periods in sequence is very strong. Moreover, by inspecting the predictions of these models we saw that the mean length of the sentences in the predicted transcriptions are very close to the ones in the training datasets (difference less than 1 most of the time). Although its number of parameters is slightly higher than that of the model RNN, the use of attention and CRF end up helping in the quality of the transcriptions. By looking at the results for Experiments 2 and 3, we also chose the RNN model, trained with Cinderella + Dog + Lucia datasets and with all classes, to be used as a unique and generic sentence segmentation detector due to its generalization performance.

7. Conclusions and Future Work

In this paper, our main goal was to develop a robust and generic sentence segmentation system for narratives of language tests, based on experiments using four datasets of narratives used to evaluate cognitive processes. Instead, our study allowed us to develop and choose two new models — RNN and RNN + ATTN + CRF — for segmenting impaired speech transcriptions, along with an ideal combination of datasets and specific groups of narratives to be used as the training set. We chose the RNN + ATTN + CRF model for creating a segmentation detector for a specific task, because it returned the best results in Experiments 1 and 2. By analyzing the results from the Experiment 3, we chose the RNN for creating our generic segmentation detector. These findings are consistent with the model selected as the one that generalizes better for a different stimuli in Experiments 2 and 3. We also made publicly available the four datasets used in this study. Although we got better segmentations by applying Viterbi decoding for all of our models with CRF on top, the input for the Viterbi algorithm is the entire transcription, and therefore a global optimization over the sequence is being done, which might not be helpful at the end because there are several valid combina-

Model	Training datasets	Test dataset	Training class	MCIs	Controls	ADs	Avg. 1	Avg. 2
1. RCNN	Cinderella + Lucia + Wallet	Dog	Same class	0.49	0.59	0.61	0.56	0.60
			All classes	0.59	0.66	0.60	0.62	**0.65**
			MCIs and Controls	0.55	0.55	0.56	0.55	**0.65**
2. RNN	Cinderella + Lucia + Wallet	Dog	Same class	0.58	0.66	0.61	0.61	0.61
			All classes	0.62	0.65	0.59	0.62	**0.66**
			MCIs and Controls	0.55	0.64	0.57	0.59	0.64
3. RNN + CRF	Cinderella + Lucia + Wallet	Dog	Same class	0.54	0.56	0.56	0.56	0.58
			All classes	0.62	0.65	0.60	0.62	0.63
			MCIs and Controls	0.58	0.65	0.54	0.59	**0.64**
4. RNN + ATTN + CRF	Cinderella + Lucia + Wallet	Dog	Same class	0.53	0.57	0.50	0.53	0.56
			All classes	0.61	0.65	0.63	0.63	**0.65**
			MCIs and Controls	0.44	0.56	0.44	0.48	0.63
5. RCNN	Cinderella + Dog + Wallet	Lucia	Same class	n/a	0.71	0.56	0.63	
			All classes	n/a	0.75	0.69	0.72	
			MCIs and Controls	n/a	0.74	0.69	0.72	
6. RNN	Cinderella + Dog + Wallet	Lucia	Same class	n/a	0.71	0.56	0.64	
			All classes	n/a	0.72	0.62	0.67	
			MCIs and Controls	n/a	0.71	0.70	0.71	
7. RNN + CRF	Cinderella + Dog + Wallet	Lucia	Same class	n/a	0.76	0.59	0.67	
			All classes	n/a	0.72	0.63	0.68	
			MCIs and Controls	n/a	0.75	0.69	0.72	
8. RNN + ATTN + CRF	Cinderella + Dog + Wallet	Lucia	Same class	n/a	0.74	0.51	0.62	
			All classes	n/a	0.76	0.62	0.69	
			MCIs and Controls	n/a	0.77	0.71	**0.74**	
9. RCNN	Cinderella + Dog + Lucia	Wallet	Same class	0.63	0.71	n/a	0.67	
			All classes	0.70	0.74	n/a	0.72	
			MCIs and Controls	0.73	0.76	n/a	**0.74**	
10. RNN	Cinderella + Dog + Lucia	Wallet	Same class	0.64	0.70	n/a	0.67	
			All classes	0.75	0.76	n/a	**0.75**	
			MCIs and Controls	0.71	0.71	n/a	0.71	
11. RNN + CRF	Cinderella + Dog + Lucia	Wallet	Same class	0.55	0.60	n/a	0.58	
			All classes	0.69	0.65	n/a	0.67	
			MCIs and Controls	0.70	0.67	n/a	0.69	
12. RNN + ATTN + CRF	Cinderella + Dog + Lucia	Wallet	Same class	0.61	0.63	n/a	0.62	
			All classes	0.72	0.67	n/a	0.69	
			MCIs and Controls	0.73	0.70	n/a	0.72	
13. RCNN	Wallet + Dog + Lucia	Cinderella	Same class	0.57	0.60	0.49	0.55	
			All classes	0.57	0.59	0.52	0.56	
			MCIs and Controls	0.60	0.61	0.52	0.58	
14. RNN	Wallet + Dog + Lucia	Cinderella	Same class	0.54	0.59	0.47	0.54	
			All classes	0.60	0.61	0.53	0.58	
			MCIs and Controls	0.60	0.61	0.51	0.57	
15. RNN + CRF	Wallet + Dog + Lucia	Cinderella	Same class	0.56	0.60	0.36	0.51	
			All classes	0.59	0.61	0.47	0.55	
			MCIs and Controls	0.61	0.63	0.50	0.58	
16. RNN + ATTN + CRF	Wallet + Dog + Lucia	Cinderella	Same class	0.53	0.60	0.26	0.46	
			All classes	0.60	0.61	0.51	0.57	
			MCIs and Controls	0.61	0.62	0.51	0.58	

Table 7: F_1 scores of the 16 models trained on a combination of three datasets and tested on the fourth remaining dataset. "Same class" means that the method was trained only with the same specific class used for testing. "All classes" means that data from all classes were used for training. Avg. 1 means average of F_1 in a row; Avg. 2 means average of F_1 by model and training class, averaging over all the combinations of training/test datasets.

tions of boundary and non-boundary states. Thus, a local decoding strategy, like posterior decoding, might be more helpful in this scenario. Another direction is to follow the recent trend of the NLP community and encode our input using contextual representations from pretrained language models like ELMo and BERT (Peters et al., 2018; Devlin et al., 2019), yet a fine-tunning procedure on a large dataset of spontaneous speech transcriptions is still probably needed (Howard and Ruder, 2018). Finally, we plan to do evaluations with the output of an ASR system, as a high word recognition error rate can greatly affect our results.

8. Bibliographical References

Aluisio, S., Cunha, A., and Scarton, C. (2016). Evaluating progression of Alzheimer's disease by regression and classification methods in a narrative language test in Portuguese. In *Proceedings of 12th International Conference on Computational Processing of the Portuguese Language*, pages 109–114, Tomar, Portugal. Springer International Publishing.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Batista, F., Moniz, H., Trancoso, I., and Mamede, N. (2012). Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. *IEEE Transactions on Audio, Speech, and Language Processing*, pages 474–485.

Bayles, K. and Tomoeda, C. (1993). *ABCD: Arizona Battery for Communication Disorders of Dementia*. Tucson, AZ: Canyonlands Publishing.

Beltrami, D., Gagliardi, G., Rossini Favretti, R., Ghidoni, E., Tamburini, F., and Calzà, L. (2018). Speech analysis by natural language processing techniques: A possible tool for very early detection of cognitive decline? *Frontiers in Aging Neuroscience*, 10(369), Nov.

Borges dos Santos, L. and Aluísio, S. (2020). Identificação automática de unidades de informaçao em testes de reconto de narrativas usando métodos de similaridade semantica. *Linguamática*, 11(2):47–63, Jan.

Bradbury, J., Merity, S., Xiong, C., and Socher, R. (2016). Quasi-recurrent neural networks. *CoRR*, abs/1611.01576.

Che, X., Wang, C., Yang, H., and Meinel, C. (2016). Punctuation prediction for unsegmented transcript based on word vector. *LREC*, pages 654–658.

Christensen, H., Gotoh, Y., and Renals, S. (2006). Punctuation annotation using statistical prosody models. *ISCA Tutorial and Research*.

da Cunha, A. L. V., de Sousa, L. B., Mansur, L. L., and Aluísio, S. M. (2015). Automatic proposition extraction from dependency trees: Helping early prediction of alzheimer's disease from narratives. *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, pages 127–130.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Drummond, C., Coutinho, G., Fonseca, R. P., Assunção, N., Teldeschi, A., de Oliveira-Souza, R., Moll, J., Tovar-Moll, F., and Mattos, P. (2015). Deficits in narrative discourse elicited by visual stimuli are already present in patients with mild cognitive impairment. *Frontiers in Aging Neuroscience*, 7(96), May.

Filiou, R.-P., Bier, N., Slegers, A., Houzé, B., Belchior, P., and Brambati, S. M. (2019). Connected speech assessment in the early detection of alzheimer's disease and mild cognitive impairment: a scoping reviews. *Aphasiology*, Apr.

Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., and Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55:43–60. Language, Computers and Cognitive Neuroscience.

Fraser, K. C., Ben-David, N., Hirst, G., Graham, N., and Rochon, E. (2015). Sentence segmentation of aphasic speech. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 862–871, Denver, Colorado, May–June. Association for Computational Linguistics.

Fraser, K. C., Rudzicz, F., and Hirst, G. (2016). Detecting late-life depression in Alzheimer's disease through analysis of speech and language. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 1–11, San Diego, CA, USA, June. Association for Computational Linguistics.

González-Gallardo, C.-E. and Torres-Moreno, J.-M. (2018). Sentence boundary detection for french with subword-level information vectors and convolutional neural networks. *ArXiv*.

Goodglass, H., Kaplan, E., and Barresi, B. (1983). *The Assessment of Aphasia and Related Disorders*. The Assessment of Aphasia and Related Disorders. Lippincott Williams & Wilkins.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Huang, G., Xu, C., Xiao, X., Xie, L., Chng, E. S., and Li, H. (2014). Multi-view features in a dnn-crf model for improved sentence unit detection on english broadcast news. In *APSIPA*.

Hübner, L. C., Loureiro, F., Tessaro, B., Siqueira, E. C. G., Jerônimo, G. M., and Smidarle, A. (2019). Bale: Bateria de avaliação da linguagem no envelhecimento. In Nicolle Zimmermann, et al., editors, *Tarefas de avaliação neuropsicológica para adultos: memória e linguagem*, volume 3. Memnon, Rio de Janeiro, 1 edition.

Khomitsevich, O., Chistikov, P., Krivosheeva, T., Epimakhova, N., and Chernykh, I. (2015). Combining prosodic and lexical classifiers for two-pass punctuation detection in a russian asr system. pages 161–169.

Klejch, O., Bell, P., and Renals, S. (2016). Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches. In *IWSLT*.

Klejch, O., Bell, P., and Renals, S. (2017). Sequence-to-sequence models for punctuated transcription combing lexical and acoustic features. In *ICASSP*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Le Boeuf, C. (1976). *Raconte: 55 historiettes en images*. L'École.

Liu, Y., Chawla, N. V., Harper, M. P., Shriberg, E., and Stolcke, A. (2005b). A study in machine learning from im-

balanced data for sentence boundary detection in speech. *Computer Speech and Language*.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, pages 127–154.

Shriberg, E., Favre, B., Fung, J., Hakkani-tür, D., and Cuendet, S. (2009). Prosodic similarities of dialog act boundaries across speaking styles. *Linguistic Patterns in Spontaneous Speech*, pages 213–239.

Tilk, O. and Alumäe, T. (2016). Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Interspeech*, pages 3047–3051.

Tillas, A. (2015). Language as grist to the mill of cognition. *Cognitive Processing*, 16(3):219–243, Aug.

Toledo, C. M., Aluísio, S. M., Dos Santos, L. B., Brucki, S., Trés, E. S., d. O. M. O., and Mansur, L. L. (2017). Analysis of macrolinguistic aspects of narratives from individuals with alzheimer's disease, mild cognitive impairment, and no cognitive impairment. *Alzheimer's dementia (Amsterdam, Netherlands)*, 10.

Treviso, M., Shulby, C., and Aluísio, S. (2017a). Evaluating word embeddings for sentence boundary detection in speech transcripts. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 151–160, Uberlândia, Brazil, October. Sociedade Brasileira de Computação.

Treviso, M., Shulby, C., and Aluísio, S. (2017b). Sentence segmentation in narrative transcripts from neuropsychological tests using recurrent convolutional neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 315–325, Valencia, Spain, April. Association for Computational Linguistics.

Treviso, M. V., dos Santos, L. B., Shulby, C., Hübner, L. C., Mansur, L. L., and Aluísio, S. M. (2018). Detecting mild cognitive impairment in narratives in brazilian portuguese: first steps towards a fully automated system. *Letras de Hoje*, 53(1):48–58, jan.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wechsler, D. (1997). *Wechsler Memory Scale - Third Edition.* The Psychological Corporation, San Antonio, TX.

Yancheva, M., Fraser, K., and Rudzicz, F. (2015). Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pages 134–139, Dresden, Germany, September. Association for Computational Linguistics.