Automatic Quantitative Prediction of Severity in Fluent Aphasia Using Sentence Representation Similarity

Katherine Ann Dunfield, Günter Neumann
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
{katherine.dunfield, guenter.neumann}@dfki.de

Abstract

Aphasia is a neurological language disorder that can severely impair a person's language production or comprehension abilities. Due to the nature of impaired comprehension, as well as the lack of substantial annotated data of aphasic speech, quantitative measures of comprehension ability in aphasic individuals are not easily obtained directly from speech. Thus, the severity of some fluent aphasia types has remained difficult to automatically assess. We investigate six proposed features to capture symptoms of fluent aphasia — three of which are focused on aspects of impaired comprehension ability, and evaluate them on their ability to model aphasia severity. To combat the issue of data sparsity, we exploit the dissimilarity between aphasic and healthy speech by leveraging word and sentence representations from a large corpus of non-aphasic speech, with the hypothesis that conversational dialogue contains implicit signifiers of comprehension. We compare results obtained using different regression models, and present proposed feature sets which correlate (best Pearson $\rho = 0.619$) with Western Aphasia Battery-Revised Aphasia Quotient (WAB-R AQ). Our experiments further demonstrate that we can achieve an improvement over a baseline through the addition of the proposed features for both WAB-R AQ prediction and Auditory-Verbal Comprehension WAB sub-test score prediction.

Keywords: Aphasia, Quantitative Severity Prediction, Comprehension Impairment, Low Resource

1. Introduction

Aphasia is a neurological language disorder, often resulting from stroke, that is characterized by language impairments that affect the production or comprehension of spoken language. Although studies have found that frequent and intensive post-stroke rehabilitation for aphasia is most beneficial in the acute stage following a stroke (Laska et al., 2011; Bhogal et al., 2003), persons with aphasia (PWA) are not always able to obtain the intensity of treatment they need during this stage, or even in later chronic stages. Depending on the location and size of the brain damage acquired, aphasia type and severity can be incredibly variable, where a PWA may exhibit a wide range of language deficits and symptoms. These variations can make it difficult to uniformly extract features of aphasia, particularly symptoms that are not explicitly expressed in a PWA's speech, such as comprehension impairments. Nonetheless, accurate predictive modelling of aphasia severity offers possibilities in facilitating more personalized and intensive treatment for aphasic patients.

Within the field of natural language processing, considerable previous work has been done in both detecting aphasia and adapting existing technology to be of better used by PWAs (Adams et al., 2017; Le et al., 2017; Fraser et al., 2014a; Fraser et al., 2014b; Thomas et al., 2005; Fraser et al., 2014c). However, due to the differences in nature of aphasia types, the primary focus of this research has been on non-fluent aphasias, which are distinguished predominately by observable production errors, and are therefore easier to obtain from narrative elicitations. Fluent aphasia, on the other hand, especially fluent aphasias noted by impairments in comprehension and semantically incoherent speech, are more difficult to observe outside of a conversational setting where confirmation of whether a lapse in comprehension has occurred can be established.

Quality data for aphasia speech is rather limited, since it takes comparatively more time and effort to find and record post-stroke aphasic speech than it does for other types of spoken language data. Likewise, because data regarding aphasia deals with real people and often needs to include significant real-life data to be useful, privacy issues become a major concern, as is often the case in medical data. The basis of the approach in extracting features for aphasia without significant training data is to leverage the dissimilarity of aphasic speech with abundant non-aphasic training data, using a few proposed methods. With the assumption that the non-aphasic data offers a survey of healthy speech, deviation from this speech can be viewed as a symptom of aphasia. By using non-aphasic speech as a baseline and computing features through dissimilarity, we create an approach that does not rely on sizeable training data of aphasic speech.

A defining characteristic of many fluent-aphasia types is a lack of understanding of both written and auditory input. As previously mentioned, much work has been performed on identifying features suitable for non-fluent aphasia, focusing on relatively surface level features, such as word frequency and speed of speech. Fluent aphasia, on the other hand, will often differ less from non-aphasic speech than the non-fluent varieties of aphasia, and is instead characterized by a lack of semantic coherency and deteriorated comprehension abilities. Therefore, to capture comprehension impairments in conversational discourse, we assume that comprehension errors often result in inappropriate responses to comments in conversational discourse.

It can be argued that since aphasia severity is expressed in a multitude of ways, achieving reliable modelling of aphasia rehabilitation depends on the availability of data that covers a sufficient range of aphasia types and symptoms, and a method of better capturing the more implicit symptoms of aphasia. In this work, we propose an investigation into a set of features, specifically selected to capture the primarily distinctive features of fluent aphasia types. These features may be extracted using state-of-the art methods in natural language processing that allow for analysis of the semantic content of speech. We therefore present three main contributions aimed to overcome issue related to data sparsity and implicit feature extraction: a method of automatically assessing comprehension ability in conversational discourse, leveraging the dissimilarity between healthy and aphasic data to estimate the degree of severity, and utilizing a metric learning approach to capture the likelihood of an aphasic utterance as to track aphasia severity in a measurable way.

2. Related Work

Qualitative classification of aphasia types (Fraser et al., 2013b; Fraser et al., 2013a; Peintner et al., 2008; Fraser et al., 2014c; Fraser et al., 2016; Vincze et al., 2016; Bucks et al., 2000; Guinn and Habash, 2012; Meilán et al., 2014; Jarrold et al., 2014) has been the primary focus of computational research into aphasia, whether in differentiating PWA's and controls or between aphasia sub-types. Traditional features sets include features that target dysfluency, lexical diversity, syntactic deviation, and language complexity. Quantitative prediction methods focus on assessing speech-based features quantitatively with the goal of providing feedback to aphasic patients. Automatic speech recognition (ASR) systems developed for aphasic speech are used to automatically extract and align a number of feature sets (Le et al., 2018; Le et al., 2014), targeting specific suggested characteristics of Aphasia. In quantitative prediction, regression models are trained on the extracted features from a subset of the annotated aphasia data.

Information-theoretic approaches (Pakhomov et al., 2010) of using the perplexity of a trained language model have been investigated in the classification of aphasia types related to dementia. The primary contribution of this research is an n-gram statistical language model trained on speech from a healthy population and used to capture unusual words and sequences from the speech of patients with frontotemporal lobar degeneration (FTLD). This model was then used to measure the dissimilarity and degree of deviation from the healthy speech data, and found that the perplexity of a language model is sensitive to the semantic deficits in FTLD patients' speech, which is often syntactically intact but is full of statistically unexpected word sequences. The perplexity index also discriminated mild from moderate-to-severely impaired FTLD patients, meaning that it is likewise sensitive to the severity of the aphasia.

Few works, to our knowledge, attempt to model comprehension. Prud'hommeaux and Roark (2015), however, explore features based on the idea that non-aphasic individuals recounting a narrative are likely to use similar words and semantic concepts to the ones used in the narrative, and suggest that this similarity can be measured using techniques such as latent semantic analysis (LSA) or cosine distance. A key element in extracting instances of comprehension impairment is the assumption that breakdowns of language understanding within conversation result in unexpected responses to a given comment or question. As outlined by Chinaei et al. (2017), these unexpected responses may follow certain trends, such as lack of continuation of topic or requests for repetition. In Watson (1999), those with Alzheimer's Disease (AD) were most likely to respond during comprehension difficulties by either a lack of continuation (no contribution or elaboration on the topic, or complete change of topic) or reprise with dysfluency (a partial or complete repetition of the question with frequent pauses and filler words). This is in contrast to those without AD, who showed more preference for specific request for information or hypothesis formation (guessing missed information).

3. Data

Datasets containing various types of conversational language are available for use in training the methods within this approach. The main requirements being that they have a clear distinction between speakers and have some sort of turn-taking conversational flow. Effort was made to collect datasets of predominantly North American English, as the test set contains mainly North American participants or at least consists primarily of participants born in the United States. For our purposes, two datasets were source to be used separately: a dataset of aphasic language to be used as a test set with both Aphasic particiapnts and controls (AphasiaBank) on which we can assess the extracted the features, and a large non-aphasic corpus that can be used to generate statistical information and examples of presumed healthy speech (Reddit).

3.1. AphasiaBank

The primary aphasic corpus used in this research is AphasiaBank, a multimedia dataset of interactions between patients with aphasia (PWA) and research investigators, for the study of communication in aphasia (MacWhinney et al., 2011; Forbes et al., 2012). The data is collected by various research groups under varying conditions following these protocols. The basic structure of these protocols involves the research investigator asking open-ended questions to elicit spontaneous verbal responses from the patient. For example, the main AphasiaBank protocol contains questions such as "*How do you think your speech is these days*?" and "*Tell me as much of the story of Cinderella as you can*". Alternatively, there is the Scripts protocol, which is less frequent, but is used by a small subset of the data (Le, 2017). The protocols contain four different discourse tasks, such as giving personal narratives in response to questions, picture descriptions, story telling, and procedural discourse. For these activities, investigators follow a script, which includes a second level prompt if the patient does not respond in ten seconds and an additional troubleshooting script with simplified questions if the patient is still not able to respond.

The AphasiaBank dataset contains a total of 431 (255 Male, 176 Female) aphasic subjects and 214 (94 Male, 120 Female) controls, with an average age of 62.4 for the aphasic group and 58.9 for the control group. The distribution of diagnosed aphasia types is outlined in Table 1.

Aphasia Type	Gender		Total
Broca	66M	33F	99
Transmotor	5M	5F	10
Global	4M	0F	4
Wernicke	21M	9F	30
Conduction	40M	26F	66
Anomic	79M	60F	139
TransSensory	0M	2F	2
AphasicNoDiagnosis	28M	18F	46
NotAphasicByWAB	12M	23F	35

Table 1: Number of AphasiaBank participants for each type of Aphasia as classified by WAB-R AQ

Speech in AphasiaBank is transcribed using the CHAT format (MacWhinney, 2000), which includes annotation of filler words, repetition, non-verbal actions, and phonetic transcription in the International Phonetic Alphabet (IPA) of word-level errors. For the purpose of this work, annotations denoting auditory occurrences and physical movements are not retained. The text for both investigator comments and replies is pre-processed and normalized using the following procedure, where the text is first lowercased, all non-alphabetic characters and punctuation are removed, and any paraphasias or neologisms marked in the annotation are replaced with an ⟨**UNK**⟩ token. Annotated speech segments between researcher and patient are extracted to create comment-reply pairs. To extract consistent comment-reply pairs, utterances that have been split in the original data, and thus do not have a direct pair with a question or comment for an investigator, are appended to the end of the preceding utterance. The resulting textual data consists of 18,038 comment-reply pairs for the aphasia subset and a total of 448,337 words (8439 unique words) of annotated aphasic speech. The control group includes an additional 2620 comment-reply pairs, with 354,620 total words and 10,012 unique words.

3.1.1. Participant-level Assessment Statistics

The AphasiaBank data provides additional information about the participant, such as a number of test scores that aim to assess the severity and type of aphasia of each aphasic speaker. This includes the Western Aphasia Battery-Revised (WAB-R) Aphasia Quotient (AQ) (Kertesz, 2006), which is the most useful for this research. WAB-R AQ is the most widely administered test in the AphasiaBank database, and is composed of multiple standardized sub-tests that targets specific aphasia-related impairments. WAB-R AQ has been shown to be a relatively reliable assessment of aphasia severity, with it demonstrating a high retest reliability in studies of chronic aphasia patients (Kertesz and Poole, 1974). Weighted performance over a number of sub-tests produces an overall score ranging from 0 to 100, that measures a speaker's general linguistic abilities and severity of their aphasia (Kertesz, 2006).

The specific sub-test groups that WAB-R AQ is composed of include: Spontaneous Speech, Repetition, Naming/Word Finding, and Auditory-Verbal Comprehension. Scores over 76 can be roughly classified as mild, whereas scores below 50 and 25 can be considered as severe and very severe respectively (Le, 2017). Following the WAB-AQ scores, the distribution of aphasia severity in the data is 47.0% mild, 38.6% moderate, 10.4% severe, and 3.9% very severe. The distribution of aphasia severity assessed by WAB-R AQ for each aphasia type is presented in Figure 1. For the purpose of evaluating fluent aphasia predictions, we consider the complete WAB-R AQ, as well as the Auditory-Verbal Comprehension sub-test scores.

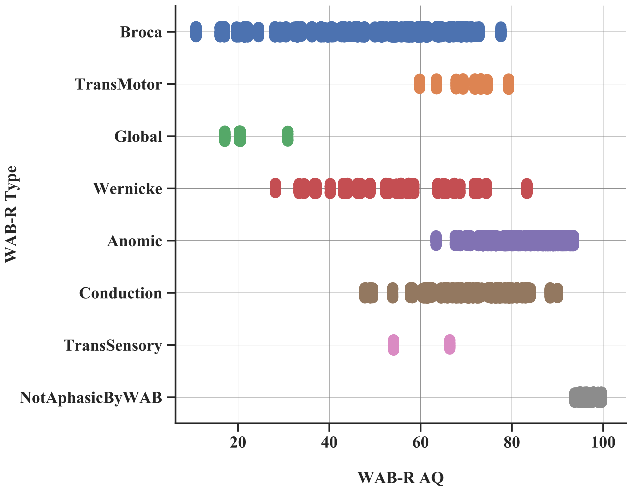

Figure 1: Distribution of WAB-R AQ Scores for each Aphasia Type (WAB-R AQ Type).

The WAB-R Auditory-Verbal Comprehension sub-test scores offer the opportunity for us to evaluate our features on whether they do accurately capture information regarding comprehension impairments and not just additional information associated with other deficits related to aphasia. Auditory-Verbal Comprehension scores are assessed by yes/no questions that may be answered in either verbal or nonverbal fashion, word recognition tasks, and by response to sequential commands, with 10.0 being the upper bounds of the test. This score is aggregated with other sub-tests as a portion of the complete WAB-R AQ. In our data, Auditory-Verbal Comprehension scores exist for 351 speakers.

3.2. Reddit

Reddit is a social news aggregation, web content rating, and discussion website with over 234 million unique users, as of March 2019. The website is primarily in English, being the 6th most visited website in the United States, and with 53.9% of its users residing in the United States, and an additional 14.5% of its user base coming from the United Kingdom and Canada (Alexa Internet, 2018). Online community forums offer an abundant source of diverse structured semi-conversational textual data to be used for training, with Reddit's being particularly easy to obtain. It should be considered semi-conversational due to the narrative quality of some comments, but there is a general assumption that threads are conversational in nature. Threads of comments are also divided hierarchically, so extracting comment relationship is possible. Though all datapoints cannot confirmed to be neurotypical or non-aphasic, the size of the dataset should minimize the impact of such outliers.

The bulk of Reddit comments dating back to its creation are obtained in JSON format from a repository prepared by (Baumgartner, 2018; Gaffney and Matias, 2018). Due to the size of the data, we only use a subset of the Reddit data. The data was naturally divided by subreddits, so a single subreddit with a still sizeable amount of data was chosen, r/IAmA (subreddits are denoted on Reddit using an r/ construction). The dialogue from this subreddit is generally representative of average healthy speech, as it is relatively serious in content, non-technical, and conversational. For normalization purposes, formatting tags are removed and double quotation marks were changed to single quotations. Links contain little relevant information for our purposes, so they were removed, along with the comments marked [**DELETED**] or [**REMOVED**]. To generate reasonable response lengths for conversation, comments longer than 50 words or 1000 characters were filtered, in addition to the removal of potential spam comments (with a user-assigned comment score $\leq$ 1). The data was further normalized to be better comparable to the other datasets used in this work. This included lowercasing the text, removing punctuation, and removing any commented links or quotations of other comments. The resulting dataset contains 1,050,699 sentence pairs, comprising of 768,348 unique words. The average number of tokens in a comment is roughly 16, where a comment is sometimes composed of multiple sentences.

4. Methods

Quantifiable measures of characteristic features of the fluent aphasia sub-type may be required to better accurately predict a general quantitative measure of aphasia severity. We propose multiple methods of extracting these measures, based on extensions of existing approaches in parallel domains, as well as additional novel approaches. We focus specifically on methods that extract features related to the production and comprehension issues found in fluent aphasic language.

4.1. Production Analysis Measures

This group of features targets aspects of fluent aphasia related to the production of language, such as sentence predictability and flow, along with occurrence of likely paraphasias or neologisms.

4.1.1. Bigram Perplexity

In previous research introduced by (Pakhomov et al., 2010), bigram language model perplexity, as well as the out-of-vocabulary (OOV) rate, of utterances were shown to have a moderate best correlation (r=0.52) with aphasia severity in dementia patients. For this approach, we suggest investigating whether we can extrapolate this approach for use with post-stroke PWAs and whether the same degree of correlation can be achieved. Following this research, we compute the probability of a sequence of words based on our non-aphasic Reddit corpus, $P(W) = P(w_1, ..., w_n)$. To compute this, we want to consider the probability of a word given its previous context, $P(w_n \,|w_{n-1})$. The probability for each bigram in our language model is computed as follows, where add-alpha smoothing is chosen and alpha α is set to 0.02, to penalize OOV words.

$$P^*(w_n \mid w_{n-1}) = \frac{C(w_{n-1}, w_n)+\alpha}{C(w_{n-1})+\alpha|V|}$$

Perplexity is then calculated for each utterance provided by the speaker, and the sum of all perplexity scores provides a speaker-level score. Perplexity in this case is measuring how well the given utterance mimics healthy speech when it comes to constructing probable strings of words.

4.1.2. Out-of-vocabulary Rate

Out-of-vocabulary (OOV) rate may reflect the rate of paraphasia or neologisms in an utterance, with neologisms in particular being characteristic in some fluent aphasias, such as Wernicke's. Often seen in the speech of patients with fluent aphasia are utterances that are long, but full of such neologisms. Therefore, a vocabulary is selected based on our non-aphasic corpus, and the target calculation would be the sum of all words not found in vocabulary over the total words in an utterance.

4.1.3. Text Imputation Similarity

Despite sounding fluent at the surface level, fluent aphasia speech often lacks semantic cohesion within an utterance. Words selected by aphasic individuals may appear semantically incongruous with other nearby words in the utterance, although some meaning may still be parsed from the utterance.

A solution to capture this aspect would be to use a language model with a much greater ngram size. However, this would require a huge corpus and rare but semantically plausible utterances would be unfairly penalized by the language model. Word embeddings in this case give much more flexibility. To describe this approach, we will consider N be the length of the input sentence, and $n = 0$ the index of the current word in the sentence. The process can then be summarised into the following steps, as shown in Figure 2, assuming a sentence string as input:

Given an utterance string S, the string is tokenized, such that $S = \{w_1, ..., w_N\}$. N copies of the input strings are created, where for each string, the $n_{th} + 1$ word is masked with the [MASK] token. Then, each masked word is predicted from the complete sentence context and resulting predicted words are concatenate to produce an output string O. The cosine similarity between the sentence vector of the original input utterance v_S and the sentence vector of the output string v_O is then compared.

4.2. Comprehension Analysis Measures

Comprehension analysis measures uniquely target fluent aphasia by assessing response predictability and sudden changes in topic, where unpredictable responses may signify a lapse in comprehension.

4.2.1. Question-Answer Similarity

Semantic relation between questions or statements and their responses are of particular interest, due to the proposed hypothesis that responses denoting an error in understanding

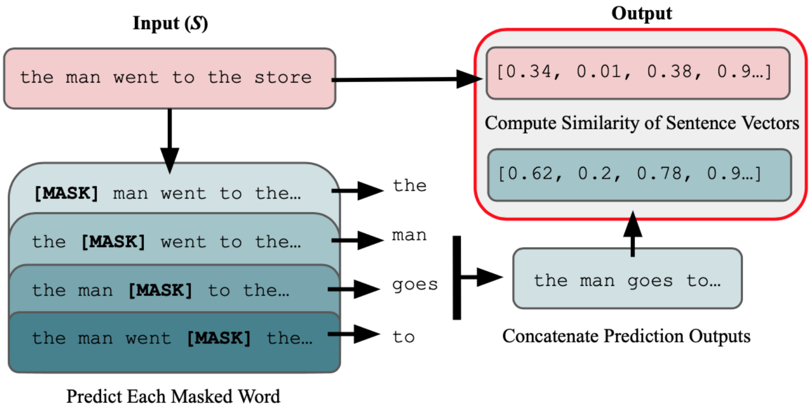

Figure 2: Text Imputation feature extraction process.

will be dissimilar to the question in words-use and semantics. To capture this, a basic measure of the cosine similarity between the sentence representations of the questions and answers in a dataset of aphasic speech can be obtained. For each question-answer pair, sentence representations of the question and answer is separately produced. The cosine similarity between the two vectors will then be computed to produce a score. Our hypothesis will be that lower similarity between the two sentence vectors will indicate less semantic overlap between the content of the sentences, meaning that the response in the question-answer pair may not be semantically coherent with the question.

For example, given a question such as '*What did you see at the zoo?*' or '*What's your favourite animal*?', answers containing few or no words related to zoo or animal, may indicate a misunderstanding of the question. The use of good sentence representations from word embedding models is especially useful in this task, because given our examples, a favourite animal might be uncommon, but still semantically related to animal.

4.2.2. Closest Question-Answer Pair

An expected and appropriate answer to a given question is assumed to closely resemble other appropriate answers to the same or similar questions. By finding the most similar question match to the question portion of a question-answer pair within a corpus of healthy speech, the question match's corresponding answer can then be compared to the answer in the input question-answer pair. Demonstrated in Figure 3, this is done by first generating the sentence representations of the input question and answer (from AphasiaBank, in our case), as well as all questions and answers in the healthy corpus (Reddit). Then, given the input question-answer sentence pair $s_{q,a}$ and a non-aphasic speech corpus of question-answer sentence pairs $C = \{c_{1,1}, ..., c_{q,a}\}$, where $q = a$. Sentence vectors for s_q and s_a are generated. For each sentence pair in C, the vector representation for c_q is also generated, resulting in a set of corpus sentence vectors VQ of $length(C)$. For each vector in VQ, its cosine similarity with s_q is computed. Selecting the vector VQ_q with greatest similarity with s_q, the sentence representation of c_a is retreived. Finally, the cosine similarity between the vectors of s_a and c_a is computed as the feature for this approach..

4.2.3. Binary Sentence Pair Classification

We leverage a binary classification approach using Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) to predict the likelihood of a given sentence pair being related. Our assumption is that question-answer sentence pairs that are predicted to be related based on our non-aphasic corpus are likely to contain semantically coherent answers to the questions, and are therefore unlikely to be characterized as a misunderstanding.

To train this approach, we first gather the non-aphasic corpora question and answer pairs collected from the Reddit dataset as positive samples and artificially fabricate negative sample pairs, by randomly sampling accompanying answers segments for each question segment from the corpus. This gives us a training sets of sentence pairs double the size of the non-aphasic corpus. With this new training set, we fine-tune a sentence pair classifier with two output classes, whether the sentences contains a valid question and answer pair or not.

The sentence pair classifier functions using the pre-trained BERT model, *bert-base-uncased*, with an additional attached classification layer. The original BERT model includes layers for language model decoding and classification, but these are not used in fine-tuning the sentence pair classifier. The sentence pair classifier uses the base model to encode the sentence representations, followed by an additional hidden, non-linear layer and the classification layer. Because the classifier uses BERT to encode the sentence representations, to fine-tune, the training data must be structured the way BERT expects, with an initial **[CLS]** token at the beginning of every sequence (question-answer pair), necessary for classification with BERT, and a **[SEP]** token between the two sentences. The classifier is then given the target question-answer pairs to generate probabilities for the two classes. The probability of the second class, which is the probability of the two sentences being a pair, is used as an aphasia severity feature.

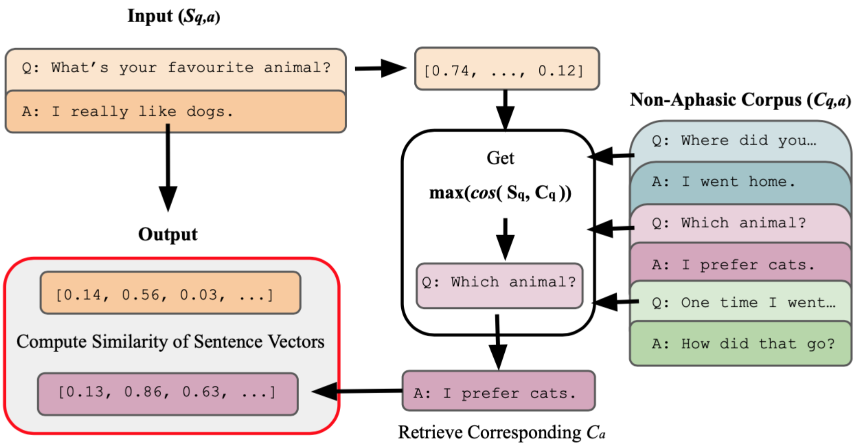

Figure 3: Closest Match feature extraction process.

5. Experimental Setup

To allow for easy comparison and combination of features that may have wildly different relationships with the data, we z-normalized all extracted features based on statistics from the control participants of AphasiaBank. Z-normalization produces a standard score useful for speaker comparison against the control group, and is calculated by subtracting the control population mean from each individual computed score and then dividing the difference by the standard deviation of the control group. With the produced feature sets, organized into groups, the goal is to produce a measure from a sample of aphasic speech that aligns with the speaker's manually diagnosed score of aphasia severity. We select only aphasic speakers who have been assigned the aphasia severity score of interest in the AphasiaBank data.

To model aphasia severity with the grouped feature sets, we use Linear Regression, Support Vector Regression (SVR), and Random Forest Regression (RFR) implemented with Scikit-learn (Pedregosa et al., 2011). The models are trained for both WAB-R AQ and Sentence Comprehension score prediction, and Pearson Correlation between the predicted results of the test set and the target scores is used to evaluate the model. The data is split at the speaker-level using four fold cross-validation, where one fourth of the data is held out as a test set during each fold, and the remaining fourths are used for training the model. While the features themselves do not require annotated aphasic data to extract, to utilize the multiple features in the most optimal way, some amount of annotated and scored aphasic data is required to fit the prediction model. We, however, also report the individual features strengths in our results.

Hyperparameter selection using 10-fold cross-validation is preformed prior to training, using the GridSearchCV function in Scikit-learn. For each model, the hyperparameters tested were:

Linear Regression Intercept $\{True, False\}$, and if intercept is calculated, then normalize $\{True, False\}$.

Support Vector Regression Penalty term C $\{1.0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$, slack parameter ϵ $\{1.0, 10^{-1}, 10^{-2}, 10^{-3}\}$, kernel type $\{rbf, linear\}$, and shrinking heuristic $\{True, False\}$.

Random Forest Regression Number of trees $\{10, 100, 200\}$, function use to measure quality of split $\{mse, mae\}$, and the max number of features to consider $\{auto, sqrt\}$.

6. Results

6.1. Individual Feature Correlations

For all feature correlations, due to likely monotonic relationships once the control data is added, we first compare the Spearman correlations of features for a combined set of aphasic and AphasiaBank control participants, in addition to the aphasic participants evaluation set, holding out the control data. Table 2 presents the correlations of all proposed features. As mentioned previously, since the control group is not naturally given a WAB-R AQ score, the scores for this group were automatically set to the upper limits of the WAB-R AQ, which is 100.0 for the full score, and 10.0 for the Auditory-Verbal Comprehension component. A feature that has a high correlation in the aphasia-only set compared to the combined control/aphasic set, likely can distinguish between more nuanced aphasia severity levels and not just between healthy controls and person with aphasia. All features have a p-value less than 0.001 in their Spearman correlations.

The feature with the strongest correlation with WAB-R AQ without the control data was the Sentence Classifier with a correlation of 0.558. This holds true also for the comprehension scores, with a correlation of 0.415. The weakest feature for the no control data is then Bigram Perplexity, likewise for both WAB-R AQ and Comprehensions with a correlation of -0.335 and -0.228 respectively. Bigram Perplexity had a very weak Pearson correlation, but still has a moderate Spearman here, indicating that it may not perform well with continuous data, but could be a useful feature in classification tasks.

Table 2: Individual Spearman correlations for all proposed features

		WAB-R AQ		Aud-Vbl Comprehension	
		With Control	**No Control**	**With Control**	**No Control**
FEATURES	Bigram Perplexity	-0.469	-0.335	-0.382	-0.228
	OOV Rate	**-0.675**	-0.465	**-0.553**	-0.251
	Text Imputation	0.505	0.372	0.395	0.273
	QA Similarity	0.281	0.345	0.221	0.271
	Closest QA Pair	0.472	0.406	0.372	0.321
	Sentence Classifier	0.33	**0.558**	0.271	**0.415**

With control data added, OOV Rate has a relatively strong correlation of -0.675. This comes with an increase of 0.21 for WAB-R AQ and 0.302 for Comprehension, compared to its correlation with the non-control data, suggesting that it may be a particularly useful feature in distinguishing healthy and aphasic individuals. Bigram Perplexity, Text Imputation, and Closest QA Pair also found a increased correlation to WAB-R when control data was added.

The Sentence Classifier feature did not correlate well with the added control data for either evaluation set, with a 0.228 difference from the non-control data. This brings it from being the most correlated feature for the non-control data to the second least with the added control data. We are unsure why this is, though we hypothesize that it is capturing variation within the control group that is not represented due to the uniform scoring the controls received. QA Similarity also correlated more strongly without the control group, though not as drastically as the Sentence Classifier.

6.2. Quantitative Aphasia Severity Prediction

One of our primary goals is to predict aphasia severity. We attempted to do so by using the features we computed in a regression model. In our preliminary investigations we utilize the three regression models for comparison. This is done on the AphasiaBank aphasic dataset with the exclusion of the AphasiaBank control group.

We compare this with a baseline consisting of a high preforming Lexical Diversity and Complexity feature set (LEX) previously used for this task (Le et al., 2018; Fraser et al., 2013b), which consists of Type-Token Ratio, a mapping of words and their frequencies in American English called the SUBTL norms (Brysbaert and New, 2009), and four additional Bristol norm word-level measures (Imageability, Age of Acquisition, Familiarity, and Phones), produced by the combined work of Stadthagen-Gonzalez and Davis (2006) and Gilhooly and Logie (1980). For our application of the baseline we achieved a Pearson correlation of 0.621 for predicting WAB-R AQ and 0.439 for predicting Auditory-Verbal Comprehension with the Support Vector Regression model. Random Forest Regression performed the best overall and for the baseline, with a correlation for 0.703 for WAB-R AQ prediction and 0.523 for Auditory-Verbal Comprehension.

We grouped our feature sets together into Production Analysis Measures (PROD), consisting of the Bigram Perplexity, OOV Rate, and Text Imputation features, and Comprehension Analysis Measures (COMP), consisting of the Question-Answer Similarity, Closest Question-Answer Pair, and Binary Sentence Pair Classification features. The proposed feature sets in the WAB-R AQ prediction task alone do not achieved the same level of results as the baseline alone, with an average correlation across models of 0.434 for Production features and 0.574 for Comprehension features. Of course each of the proposed groups consist of half of the features as the Lexical feature set. The Linear model is an exception, however, as it performs unexpectedly well with the Comprehension feature set, beating the baseline with the Comprehension features alone. Over the three models, the best performing set of features for this task is the combined baseline and the comprehension features (LEX + COMP), which given us an average correlation of 0.692 and an improvement over the baseline of 0.066. It is also the best performing feature set for both the Linear model and SVR, with Random Forest Regression performing best with all features (LEX + PROD + COMP). The Linear model performed overall, surprisingly well for the task, yielding a slightly stronger correlation than Support Vector Regression.

Predictions for Auditory-Verbal Comprehension scores follow a similar pattern to the WAB-R AQ task. The Lexical and Comprehension feature set (LEX + COMP) prediction correlations remain the best performing with an average correlation of 0.490 and an improvement over the baseline of 0.037. In the prediction of Comprehension scores, the Comprehension feature set generally performed more closely to the baseline than in the WAB-R task, whereas the Production feature set performed equally as poorly compared to the baseline as it did in predicting WAB-R AQ.

It is interesting to note that any inclusion of the Production feature set in both tasks worsened the performance of the model, with the exception of Random Forest Regression, which had the best results with the Lexical and Production features sets (LEX + PROD). This suggests that some sort of feature selection may need to be applied.

6.2.1. Feature Selection

Certain particularities stand out in the model prediction results which leaves additional consideration to the efficacy of some features, such as the decrease in improvement following the addition of the Production feature set (or Comprehension feature sets for Random Forest Regression) and the poor linear correlations of some features. For this reason, we apply a feature selection method to the data.

Table 3: Prediction model results for 3 feature sets after applying feature selection, on the two evaluations sets: WAB-R AQ and Auditory-Verbal (Aud-Vbl) Comprehension.

	Pearson *r* (p-value)		
Feature Sets	**Linear**	**Support Vector**	**Random Forest**
WAB-R Baseline	0.563	0.617	0.691
WAB-R Proposed	0.616	0.619	0.576
WAB-R Combined	0.715	0.714	0.74
Aud-Vbl Baseline	0.403	0.428	0.494
Aud-Vbl Proposed	0.414 (0.001)	0.423	0.321 (0.01)
Aud-Vbl Combined	0.488	0.491	0.537

We apply the Boruta algorithm, using Boruta_py and Scikit-learn, to optimize prediction results and as an easily interpretable method for feature selection. With this we can determine which combination of features yield the best performance from our models. The Boruta algorithm (Kursa et al., 2010) is a recursive feature elimination method. It functions by adding randomness to the data in creating shuffled copies of all the features. Then we give this extended feature set to be fit to the evaluation data using a Random Forest Regressor. Feature importance is measured during training of the regressor, using Mean Decrease Accuracy, where higher means indicate more importance. For each iteration of training, the algorithm checks if a feature has a higher importance than the best of its shuffled copies and removes features it deems as unimportant.

For each evaluation measure (WAB-R AQ Aud-Vbl Comprehension), we run feature selection on three sets of the features, one including the only baseline Lexical features, one with only our proposed feature, and one with all features. We report the model results in Table 3. P-values for all feature set predictions were less than 0.001, unless otherwise specified. For WAB-R AQ prediction, the following proposed features were selected: OOV Rate, Text Imputation, Closest QA Pair, Sentence Classifier

For Auditory-Verbal Comprehension prediction, only Closest QA Pair and the Binary Sentence Classifier probabilities were selected, both with the baseline features and without. For Auditory-Verbal Comprehension, production based features in particular were excluded during selection, such as Phone Length in the Lexical features set, and OOV Rate. In all cases the Combined Baseline and Proposed feature set performed best, though the results using proposed selected features alone correlated better than the baseline feature in all models except Random Forest Regression for WAB-R AQ Prediction. On the other hand, Random Forest Regression provided the best results using all features for both WAB-R AQ and Auditory-Verbal Comprehension score prediction.

7. Conclusion

In this work, we proposed methods for extracting six features we hypothesized would be useful in modelling symptoms consequent of fluent aphasia, such as comprehension impairments, semantic incoherence, and increased likelihood of paraphasias and neologisms. We make primary use of word and sentence representation to better assess these aspects. Our chosen approach utilized the perceived dissimilarity between aphasic and non-aphasic speech and thus did not require any annotated data of aphasic speech to obtain the proposed features. We assess the performance of our features by investigating how they benefit the task of quantitative aphasia severity prediction. Framing the task as a regression problem, and given a set of data with manually assigned aphasia severity scores, we evaluated the linear correlation of the predicted scores using our proposed features against the gold-standard severity scores. We compared these results to a baseline based on work by Le et al. (2018; Fraser et al. (2013b). Most of the proposed features alone were found to have moderate correlation with the evaluation scores, and after applying feature selection, the proposed features performed better or equal to the baseline in the regression task using Linear Regression and Support Vector Regression. For all regression models, the combined baseline and proposed features yielded the best results in all evaluation cases. Specifically, we found that the task benefits most from the inclusion of BERT sentence representations fine-tuned on a large amount of conversational data.

This work has also raised a number of questions and possible avenues for future work in this research area. Since scores were predicted at the utterance-level and then averaged, a wider range of statistics for the proposed features may yield better results, as was previously investigated by (Le et al., 2018). Likewise, given a larger dataset of aphasic language for each aphasia subtype, variations between subtypes could offer further structured results that highlight the difference between fluent and non-fluent aphasia. The practical applications of such a task using more robust feature sets, automatic speech recognition, and utterance-level assessment is also worth consideration.

8. Acknowledgements

This work was partially funded by the European Union's Horizon 2020 grant agreement No. 777107 (Precise4Q) and by the BMBF project DeepLee ((01IW17001).

9. Bibliographical References

Adams, J., Bedrick, S., Fergadiotis, G., Gorman, K., and van Santen, J. (2017). Target word prediction and paraphasia classification in spoken discourse. In *BioNLP 2017*, pages 1–8.

Alexa Internet. (2018). Reddit.com site info. [Online; accessed 2018-02-28].

Baumgartner, J. (2018). [dataset] subreddit data for over 750,000 reddit subreddits. [Online; accessed 2019-05-21].

Bhogal, S. K., Teasell, R., Speechley, M., and Albert, M. (2003). Intensity of aphasia therapy, impact on recovery. *Stroke-a Journal of Cerebral Circulation*, 34(4):987–991.

Brysbaert, M. and New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Bucks, R. S., Singh, S., Cuerden, J. M., and Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91.

Chinaei, H., Currie, L. C., Danks, A., Lin, H., Mehta, T., and Rudzicz, F. (2017). Identifying and avoiding confusion in dialogue with people with alzheimer's disease. *Computational Linguistics*, 43(2):377–406.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Forbes, M. M., Fromm, D., and MacWhinney, B. (2012). Aphasiabank: A resource for clinicians. In *Seminars in speech and language*, volume 33, pages 217–222. Thieme Medical Publishers.

Fraser, K., Rudzicz, F., Graham, N., and Rochon, E. (2013a). Automatic speech recognition in the diagnosis of primary progressive aphasia. In *Proceedings of the fourth workshop on speech and language processing for assistive technologies*, pages 47–54.

Fraser, K. C., Rudzicz, F., and Rochon, E. (2013b). Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *INTERSPEECH*, pages 2177–2181.

Fraser, K. C., Hirst, G., Graham, N. L., Meltzer, J. A., Black, S. E., and Rochon, E. (2014a). Comparison of different feature sets for identification of variants in progressive aphasia. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 17–26.

Fraser, K. C., Hirst, G., Meltzer, J. A., Mack, J. E., and Thompson, C. K. (2014b). Using statistical parsing to detect agrammatic aphasia. *Proceedings of BioNLP 2014*, pages 134–142.

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

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Gaffney, D. and Matias, J. N. (2018). Caveat emptor, computational social science: Large-scale missing data in a widely-published reddit corpus. *PloS one*, 13(7):e0200162.

Gilhooly, K. J. and Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior research methods & instrumentation*, 12(4):395–427.

Guinn, C. I. and Habash, A. (2012). Language analysis of speakers with dementia of the alzheimer's type. In *2012 AAAI Fall Symposium Series*.

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., and Ogar, J. (2014). Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 27–37.

Kertesz, A. and Poole, E. (1974). The aphasia quotient: the taxonomic approach to measurement of aphasic disability. *Canadian Journal of Neurological Sciences*, 1(1):7–16.

Kertesz, A. (2006). Western aphasia battery-revised (wab-r): Harcourt assessment. *San Antonio, TX*.

Kursa, M. B., Jankowski, A., and Rudnicki, W. R. (2010). Boruta - a system for feature selection. *Fundam. Inf.*, 101(4):271–285, December.

Laska, A., Kahan, T., Hellblom, A., Murray, V., and Von Arbin, M. (2011). A randomized controlled trial on very early speech and language therapy in acute stroke patients with aphasia. *Cerebrovascular diseases extra*, 1(1):66–74.

Le, D., Licata, K., Mercado, E., Persad, C., and Provost, E. M. (2014). Automatic analysis of speech quality for aphasia treatment. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4853–4857. IEEE.

Le, D., Licata, K., and Provost, E. M. (2017). Automatic paraphasia detection from aphasic speech: A preliminary study. In *Interspeech*, pages 294–298.

Le, D., Licata, K., and Provost, E. M. (2018). Automatic quantitative analysis of spontaneous aphasic speech. *Speech Communication*, 100:1–12.

Le, D. (2017). *Towards Automatic Speech-Language Assessment for Aphasia Rehabilitation*. Ph.D. thesis.

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

MacWhinney, B. (2000). The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database.

Meilán, J. J. G., Martínez-Sánchez, F., Carro, J., López, D. E., Millian-Morell, L., and Arana, J. M. (2014). Speech in alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, 37(5-6):327–334.

Pakhomov, S. V., Smith, G. E., Marino, S., Birnbaum, A., Graff-Radford, N., Caselli, R., Boeve, B., and Knopman, D. S. (2010). A computerized technique to assess language use patterns in patients with frontotemporal dementia. *Journal of neurolinguistics*, 23(2):127–144.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peintner, B., Jarrold, W., Vergyri, D., Richey, C., Tempini, M. L. G., and Ogar, J. (2008). Learning diagnostic models using speech and language measures. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4648–4651. IEEE.

Prud'hommeaux, E. and Roark, B. (2015). Graph-based word alignment for clinical language evaluation. *Computational Linguistics*, 41(4):549–578.

Stadthagen-Gonzalez, H. and Davis, C. J. (2006). The bristol norms for age of acquisition, imageability, and familiarity. *Behavior research methods*, 38(4):598–605.

Thomas, C., Keselj, V., Cercone, N., Rockwood, K., and Asp, E. (2005). Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In *IEEE International Conference Mechatronics and Automation, 2005*, volume 3, pages 1569–1574. IEEE.

Vincze, V., Gosztolya, G., Tóth, L., Hoffmann, I., Szatlóczki, G., Bánréti, Z., Pákáski, M., and Kálmán, J. (2016). Detecting mild cognitive impairment by exploiting linguistic information from transcripts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 181–187.

Watson, C. M. (1999). An analysis of trouble and repair in the natural conversations of people with dementia of the alzheimer's type. *Aphasiology*, 13(3):195–218.