Storyline-Centric Detection of Aphasia and Dysarthria in Stroke Patient Transcripts

Peiqi Sui[1], **Kelvin K. Wong**[1], **Xiaohui Yu**[1], **John J. Volpi**[2], and **Stephen T. C. Wong**[1]
[1] T.T. and W.F. Chao Center for BRAIN & Houston Methodist Neal Cancer Center, Houston Methodist Hospital
[2] Eddy Scurlock Comprehensive Stroke Center, Department of Neurology, Houston Methodist Hospital and Weil College of Medicine
psui@houstonmethodist.org

Abstract

Aphasia and dysarthria are both common symptoms of stroke, affecting around 30% and 50% of acute ischemic stroke patients. In this paper, we propose a storyline-centric approach to detect aphasia and dysarthria in acute stroke patients using transcribed picture descriptions alone. Our pipeline enriches the training set with healthy data to address the lack of acute stroke patient data and utilizes knowledge distillation to significantly improve upon a document classification baseline, achieving an AUC of 0.814 (aphasia) and 0.764 (dysarthria) on a patient-only validation set.

1 Introduction

Published studies reported that about 30% of acute ischemic stroke patients are presented with aphasia as an initial symptom (Engelter et al., 2006), while around half of these patients exhibit some form of dysarthria (Urban et. al., 2001), with acute dysarthria specifically associated with small lacunar stroke primarily due to small vessel disease. The "cookie theft" picture description task shown in Figure 1 is commonly used for language assessment in the NIH stroke scale (NIHSS) to score the severity of aphasia and dysarthria among others. Currently, the scoring is done by a certified healthcare worker.

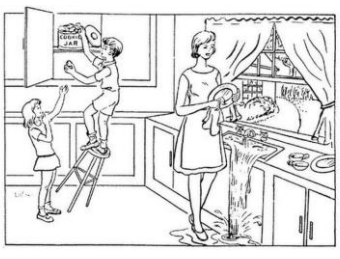

Figure 1. Cookie theft picture commonly used in patient description tasks for stroke assessment and aphasia and dysarthria diagnosis.

Recent research has demonstrated the feasibility of deep learning-based stroke detection, using facial expression and voice data gathered from the "cookie theft" storytelling task (Figure 1) that serves to differentiate mild/moderate stroke among stroke mimics in the emergency room (Cai et al., 2022). However, existing approaches to AI-enabled stroke prediction have only utilized the audio spectrum of patient recordings. The language content of these recordings is yet to be used for language assessment, even though the storytelling audio is often automatically transcribed. This motivates us to apply unstructured storytelling transcript and large-scale language model in order to predict the presence of aphasia or dysarthria in patients with stroke-like symptoms, using NIHSS subscores 9 and 10 as gold standard.

In this paper, we present a new storyline-centric pipeline that uses transcribed patient descriptions alone to detect aphasia and dysarthria in patients with stroke-like symptoms. Although no such study has been done in stroke to the best of our knowledge, these unlabeled patient transcripts are proven to be highly useful for language-related symptoms detection by a robust body of research in Alzheimer's disease (AD) prediction and monitoring (de la Fuente Garcia et al., 2020). Working with patient descriptions of the cookie theft picture amongst other transcripts, recent studies in AD discover that transformer-based language models that leverage a comprehensive language understanding (Guo et al., 2019; Qiao et al., 2021; Liu et al., 2021; Wang et al., 2023) tend to outperform models trained on syntactic, lexical, or pragmatic features alone (Fors et al, 2018; Ammar and Ayed, 2018). Moreover, models that depend on syntax and pre-defined lexicons are more prone to racial and educational biases that

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 422–432

discriminate against patients who are non-native speakers or dialect users of English.

The relative lack of NLP-enabled stroke detection could be due to the lack of patient textual data, while similar studies in Alzheimer's disease could benefit from publicly available corpora of patient narratives, both on the cookie theft picture and otherwise (MacWhinney 2019). To tackle this challenge, we are interested in exploring how data from cloud-sourced healthy volunteers, which is easier and more cost-effective to obtain, could be used to improve clinical NLP models. We experiment with two approaches to enrich our training set with healthy subject data, by including them first directly as the texts themselves and then indirectly as metadata representations in the form of knowledge graphs. By circumventing the data bottleneck, we believe that it is possible to improve NLP-enabled detection of language symptoms in stroke patients.

Major Contributions. In this work, we present 1) a pair of ELECTRA-based models for detecting aphasia and dysarthria in patient documents by performing data distillation with storyline-encoded knowledge graphs extracted from both healthy and patient transcripts, 2) de-noised document-level knowledge graphs that represent the "correct" storyline as a consensus between healthy volunteers, which provides semantic emphasis that enriches document classification, and 3) a qualitative evaluation of our models' performance that examines their semantic and clinical limitations with error-based behavioral testing.

2 Data Enrichment and Baselines

2.1 Patient Data

To build our dataset, patients with stroke-like symptoms from the Houston Methodist Hospital System are instructed to verbally describe the "cookie theft" image for one minute while their audio and facial video were recorded. Ground truth labels for aphasia and dysarthria are respectively obtained from subscores 9 and 10 of the NIH stroke assessment. The voice recordings of patients describing the image in English are automatically transcribed with Assembly AI, resulting in a dataset of 268 patient transcripts (3 patient samples are dropped due to poor quality). We retrieve subscores 9 and scale 10 scores for aphasia and dysarthria respectively from these patient transcriptions (49/268 for aphasia, 74/268 for dysarthria). Notably, 44/49 of our aphasia patients and 60/74 of our dysarthria patients are diagnosed with stroke.

2.2 Data Enrichment with Crowdsourced Healthy Volunteer Transcripts

Data enrichment refers to the process of supplementing internal data with external data sources (Allen and Cervo, 2015). In the clinical domain, it could be applied to address the lack of available patient data by enlarging the training set with healthy subjects as negative labels. We leverage Amazon's Mechanical Turk (mTurk) to collect healthy volunteer voice data from native English speakers from the United States describing the same cookie-jar theft story (n=988). We conduct manual quality control and confidence score evaluations to filter the mTurk dataset (n=675) to ensure that it only consists of high-quality audio recordings and storylines that resemble that of healthy subjects.

The healthy subject transcripts are then used to enrich the patient transcripts, and both healthy and patient transcripts were separately split at an 80/20 ratio into an enriched training set (n=754, 214 patients), which utilizes both healthy and patient data for training, and a patient-only validation set (n=54), with the proportion of each label class (aphasia or dysarthria) preserved. We exclude all healthy data points from the validation set, to make sure that measurement metrics in upcoming sections would represent the classification performance on acute stroke patients alone. We choose to not include a hold-out test set due to the lack of patient data.

2.3 Enriched Baselines for Patient Document Classification

Transformer-based methods (Vaswani et al., 2017) have been credited with most recent progress in the area of text classification (Minaee et al., 2021). We experiment with fine-tuning various transformer-based language models, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), and ELECTRA (Clark et al., 2019), to implement binary classification on our patient documents for both aphasia and dysarthria. We choose not to implement any clinical NLP models because patient descriptions of the cookie theft picture themselves are not particularly relevant to the clinical domain.

Model	BERT-base	ALBERT-large	RoBERTa	ELECTRA-large
AUC (label='Aphasia')	0.533	0.595	0.512	**0.615**
AUC (label='Dysarthria')	0.421	0.416	0.403	**0.424**
AUC (label='Combined')	0.558	0.596	0.52	**0.627**

Table 1. Baseline patient document classification performance on aphasia, dysarthria, and combined (patients with either aphasia, dysarthria, or both) labels, after data enrichment.

We first experiment with only using patient data for model training, and their performance is evaluated with the area-under-curve (AUC) metric. Due to the small size of the patient training set (n=214), the validation performance of various baseline models is unsatisfactory when trained on patient data alone, with AUC between 0.43 to 0.46. To address this, an enriched training set is created by combining healthy and patient data, while the best model is selected using a patient-only validation set.

Baseline models are established when a significant improvement in performance is achieved with data enrichment. ELECTRA-large is the best performing model overall: after enrichment, its AUC rises to 0.615 for aphasia and 0.627 for either aphasia or dysarthria in the patient-only validation set. Notably, all models' inferences on aphasia outperform that of dysarthria, as shown in Table 1. This gap in performance could be attributed to the imprecise and unintelligible speech (Yorkston, 1996) that is common in dysarthria due to poor motor coordination. As a form of language impairment, dysarthria is more often manifested as difficulty articulating rather than semantic mismatch (Mitchell et al., 2017), which might not be directly visible to language models without domain-specific fine-tuning.

Our proposed methods in the next section aim to improve upon these baseline results, as stated in the metrics of Table 1.

3 Knowledge Distillation

This section reports our experiment designs aimed at testing the hypothesis that knowledge distillation with storyline-encoded knowledge graphs, extracted from both healthy and patient transcripts, would transfer semantic knowledge to the enriched document classification model and improve the performance of detecting aphasia and dysarthria.

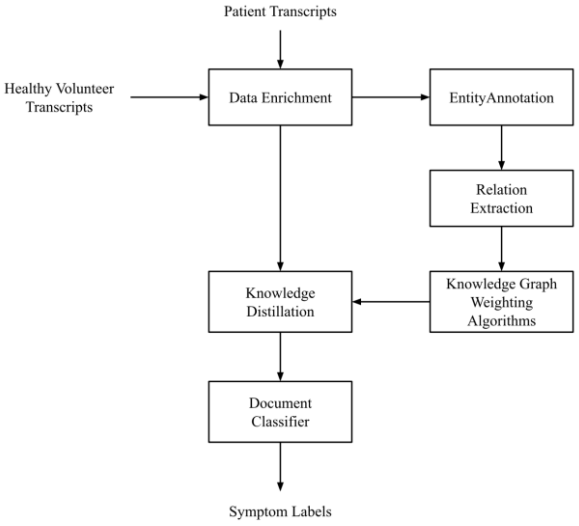

Figure 2. Schematic of the knowledge distillation pipeline.

As shown in Figure 2, the knowledge distillation pipeline has two components. First, we conduct unsupervised knowledge graph extraction with BERT-based entity annotation and relation extraction models. Second, we leverage these knowledge graphs to provide semantic reinforcement for the baseline ELECTRA document classification model defined in Section 2.3.

3.1 Knowledge Graph Extraction from Healthy Volunteer Transcripts as Representations of Ground Truth Storyline

Beyond data enrichment, we further leverage the mTurk dataset of healthy subjects to construct document-level knowledge graphs (KG) that represent the ground truth storyline of the "cookie theft" picture, which we will use to enrich the knowledge distillation learning in Section 3.3. The nodes of these KGs are key entities extracted from each transcript by a BERT-based annotator (Grootendorst, 2020), and the edges between them are semantic relations that describe a form of non-sequential storyline. For relation extraction, we implement a BERT-based model (Soares et al.,

2019)[1] fine-tuned with the general domain relation labels of SemEval-2010 Task 8 (Hedrickx et al., 2010). However, the initial outputs lack coherence and contain excessive noise. To address this, we implement the following denoising strategies sequentially:

1. **Entity permutation on the sentence level.** Each transcript is tokenized into sentences, and key phrases (n=4) are extracted on a sentence level. Each key phrase is paired with each other's key phrase and then passed into the relation extractor with the rest of the sentence. We filtered out excessively short sentences (n<=5) from going into the classifier to avoid having trivial words annotated as key phrases, and the relations classified as "Other" from going into the output knowledge graph. This results in a significant improvement from the initial approach to pass the entire paragraph into the relation extractor, which creates many contrived situations where a word at the start of the paragraph might be paired with a word at the end.
2. **Domain-specific pre-training on mTurk and patient transcripts.** We include an additional pre-training step to the relation extractor so that the language model could have some exposure to our corpus before fine-tuning and produce more relevant results.
3. **Nodes and edges cleaning.** Entities that evidently describe the same entity, i.e., "mother" and "woman," are combined. Redundant relation labels are also removed from the fine-tuning stage. The Sem-Eval 2010 Task 8 dataset's relation classification dataset contains these following relation labels that are not relevant to the cookie theft picture: "Component-Whole," "Product-Producer," "Member-Collection," and "Message-Topic." Removing them significantly reduces the number of misplaced nodes and edges in the output KGs.
4. **Updating to lighter models.** We run relation extraction on ALBERT instead of BERT, since it performed better on both the patients and enriched set during the baseline testing.

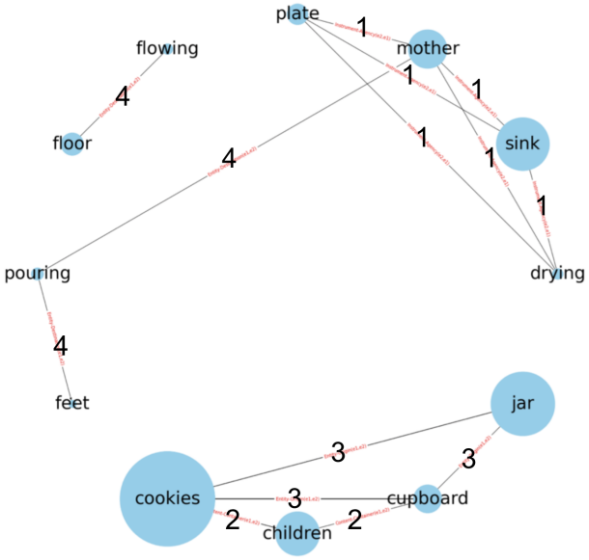

Figure 3. Example of a denoised knowledge graph. The nodes are key entities from the picture descriptions, with their size representing their phrase importance. The edges are semantic relations between the two nodes, including:

1. Instrument-Agency
2. Content-Container
3. Entity-Origin
4. Entity-Destination

An example of a typical document-level KG after denoising is shown in Figure 3. The nodes are then weighted using TF-IDF based on the collection of all entities extracted from the mTurk dataset, which indicates the relative importance of each entity.

3.2 Knowledge Distillation for Semantic Reinforcement in Aphasia and Dysarthria Document Classification

With 675 denoised KGs that represents the storyline as described by healthy subjects, we aim to conduct storyline-centric knowledge distillation learning to improve on the classification results of data enrichment alone (Section 3.1) in Table 1.

[1] We could not find the official code repository for Soares et al. (2019). Instead, we used a popular community implementation available at https://github.com/plkmo/BERT-Relation-Extraction.

Patient Passage	Extracted Triples

Okay. Oh, I've seen this before. There's a lot of things going on, and some of them aren't quite right. The **sink** is **overflowing**, and there getting some **cookies** from this cookie jar, and he's about to fall, but then she's acting like everything's okay, and it's not okay, but she must be on Prospect or something like that, because she does not even see that anything is happening over there. And he's about to **fall**, and he can have a **concussion** or something like that. And it's a nice day outside, and there's some other things in the **cabinet**, and there's nothing on this side of the **sink** or anything like that. And she's drying a dish, and there's **curtains**, and there's a walkway, and there's **sink**	sink, Cause-Effect(e2,e1), overflowing;
	cookie, Cause-Effect(e1,e2), overflowing;
	sink, Cause-Effect(e2,e1), fall;
	sink, Content-Container(e1,e2), cabinet;
	concussion, Cause-Effect(e2,e1), fall;
	curtains, Entity-Origin(e2,e1), drying;
	drying, Instrument-Agency(e2,e1), dish

Figure 4. Sample patient transcript and its extracted triples. Color scheme denotes sentence of origin, with entities bolded in the transcript.

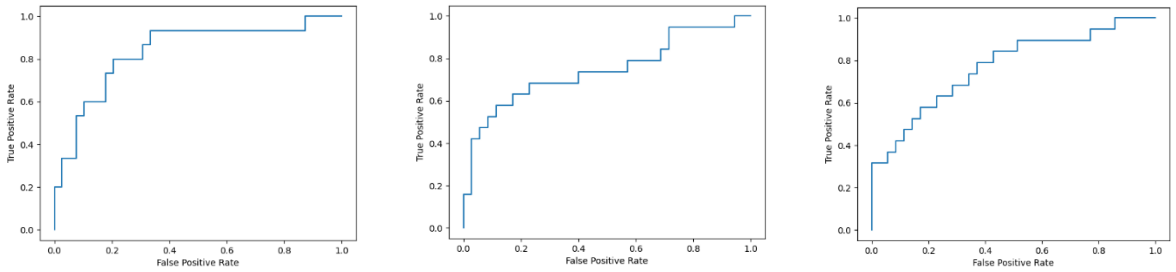

Figure 5. ROC curves of triples-distilled ELECTRA-large on aphasia (AUC=0.814), dysarthria (AUC=0.764), and combined (AUC=0.769) in patient only validation set

To achieve this, we experiment with two types of knowledge distillation: 1) triple classification, which combines all 675 KGs into one large KG, and 2) triple concatenation, which leverages the KGs individually on the document level.

To construct and extract more meaningful and accurate KGs from patient transcripts, we make the following adjustments to the methods described in Section 3.1 to increase its effectiveness in a clinical setting:

- **Joining sentences that are excessively short.** 41.8% of the patients with either aphasia or dysarthria and 37.4% of those without talk in very fragmented sentences with minimal syntax, usually in two-or-three-word sequences of "subject-verb," "subject-verb-object," or simply isolated words. To ensure the meaningfulness of the output entities, we join these sentences together to offer sufficient contexts for the BERT-based relation extractor.
- **Parameter tuning.** We evaluate the output from different key phrase counts per sentence and minimum sentence length, and qualitatively determine that four entities per sentence no shorter than 100 characters generates the best tradeoff between noise and the number of triples extracted.

A representative example of patient transcripts and their extracted triples is shown in Figure 4. While mostly accurate, the KG model is still limited by some inaccurate entity pairings as it permutates through each sentence, which we hope to address with further denoising steps outlined in Section 5.

Triple Classification: Triple classification is a KG completion task that identifies whether a triple could be a constitutive part of a certain given KG, as a naive approach to knowledge distillation. A triple is defined as a set of head, relation, and tail that is the basic constituting unit of a KG. We

implement triple classification with LMKE (Wang et al., 2022), a BERT-based model that is representative of various highly similar recent triple classification models based on the same codebase proposed by Yao et al. (2019). We collapse all patient-level KGs generated in Section 3.1 into one large KG, use it to re-train LMKE as the ground truth training set, and then evaluate the model on all patient triples as the inference set. This is in essence a zero-shot learning approach, as the training set only has positive labels. We run inference on each triple extracted from the patient transcripts to determine if they belong to the ground truth KG. The triples extracted from patients with either aphasia or dysarthria are expected to output a negative label (not belonging to the ground truth KG), since our assumption is that their descriptions do not semantically fit the "no language symptom" storyline. Unfortunately, LMKE does not facilitate patient-level prediction, while its triple-level performance (AUC=0.612, label= 'combined') does not suggest potential improvement from the baselines in Section 2.3.

Triple Concatenation: Since LMKE does not perform well on the triple level, we seek to transfer and distill the knowledge that the model learned from triple classification to the patient level. One approach we successfully implement is to concatenate triples to the end of the transcript that they are generated from as a way of data distillation for semantic reinforcement. Since the triples are directly extracted from the transcripts and the two are too correlated to be considered statistically independent, we choose to concatenate them instead of training them as two separate features. This significantly increases the performance of ELECTRA on combined language symptoms detection (AUC=0.769), and the feature dependency that motivates the concatenation is further validated by an ablation that shows using the concatenated triples alone would perform poorly (AUC=0.427). Notably, the triples-enriched model demonstrates a significant improvement on dysarthria detection (AUC=0.764), which makes its performance more balanced between different language conditions (Figure 5). Excessive repetition (Mitchell et al., 2017) is a prominent sign of dysarthria, as recent studies (Mitchell et al., 2021; Kirshner, 2022) find a repetition test to be an effective metric for dysarthria diagnosis and examination. Our use of triples in aphasia and dysarthria detection could be considered as an AI-enabled automation of the repetition test. It puts semantic emphasis on key entities that dysarthria patients struggle to articulate, which would otherwise not be visible to ELECTRA or other language models from the word embedding space alone.

4 Behavioral Testing and Discussion

We conduct further qualitative testing to thoroughly evaluate the sensitivity of our aphasia and dysarthria models to different types of language errors and generalizability to external data. This is motivated by a recent surge in the behavioral testing of NLP models that challenge the effectiveness of common quantitative testing metrics (Ribeiro et al., 2020). For NLP models in the clinical domain, van Aken et al. (2022) highlight the need to simulate plausible real-life patient inputs to analyze model sensitivity directly.

Our main goals thus are to 1) verify that our models are in fact learning semantically, as expected by our methodology, 2) simulate "external" data to assess potential model overfitting to the specific clinical settings of the patient data, and 3) better understand the semantic limitations and boundary conditions of our models in order to make more accurate, informed, and measured claims about their clinical outcomes. For each patient transcript, we generate 9 versions of the original text that amplify types of language errors in both semantic and syntactic categories. Table 2 compares the aphasia and dysarthria models' performance on all categories of error-infused transcripts, with F1 as the evaluation metric.

4.1 Syntax Testing

On the word level, we manually saturate each patient description with subject-verb disagreement, verb tense, and pronoun errors. The dysarthria model's huge drop in F1, in particular, shows that it is particularly sensitive to word-level syntax errors.

On the sentence level, we experiment with sentence structures that could cause grammatical confusion: 1) run-on sentences with too few punctuations, and 2) overly fragmented sentences with too many punctuations. 1) has been directly identified as a symptom of aphasia by the NIH's

Error Type	Example	Aphasia (F1)	Dysarthria (F1)
Original	“There’s a kid falling off a chair, trying to get a cookie. His sisters trying to take the cookie away from him. And moms washing dishes. The sinking is overflowing. She's looking out the window and the water is going all over the floor…”	0.86	0.72
Random Noise	“There’s a kidh fallibng off a chair, trying touket a cookie.qHis sisters tryingh to take the cookie iway fromq him. jAnd mems wauhing dishes. Thepslinhking is overlflowing. Shme's lokking omt thg window and the watewr gis ghing all nver the floor…”	0.62	0.33
Excessive Grammatical Errors	“There have a kid falled off a chair, tryed to get a cookie. Her sisters tries take the cookie away from her. And moms wash dishes. The sinking are overflowing. Him’s looked out the window and the water is going all over the floor…”	0.69	0.4
Run-on Sentences	“There’s a kid falling off a chair trying to get a cookie his sister’s trying to take the cookie away from him and moms washing dishes the sinking is overflowing she’s looking out the window and the water is going all over the floor…”	**0.73**	0.62
Fragmented Sentences	“There’s a. kid falling. off a chair trying. to get. a cookie. His. sisters trying. to. take. the. cookie away from him. And moms. washing dishes. The sinking is overflowing. She’s looking out. the window. and the water. is going. all over the floor…”	0.67	**0.64**
Additional Object(s)	“There’s radio a kid falling radio off a chair, trying radio to get radio a cookie. His sisters trying to take the cookie away from him. And moms washing dishes. The sinking is overflowing. She's looking out the window and the water is going all over the floor…”	0.7	0.6
Removed Key Object(s)	“There’s a kid falling off a chair, trying to get. His sisters trying to take away from him. And moms washing dishes. The sinking is overflowing. She’s looking out the window and the water is going all over the floor…”	0.67	0.54
Keeping First Sentence Only	“There's a kid falling off a chair, trying to get a cookie.”	0.7	0.25
Randomly Deleting	“to His take from dishes. The sinking window and is pretty that's about getting soaking wet. a and the boy's cookie in his right grab another one. to”	0.56	0.37
Reversed Sentence Order	“Is that no. And the sisters reaching up, trying to get one of the cookies from me. The boy’s holding cookie in his left hand as he's falling off the chair, and he's got his right hand in the cookie jar trying to grab another one. That's about all I see… There's a kid falling off a chair, trying to get a cookie.”	0.68	0.56

Table 2. Examples of cookie theft picture descriptions infused with each category of language errors for behavioral testing

most up-to-date definition[2], while 2) has been linked to dysarthria as many studies find that dysarthria patients tend to be more effective at processing shorter sentences (Allison et al., 2019), especially when aided with pauses and verbal stress-making (Kuschmann and Lowit, 2021). Both observations are supported by our results: out of all error types, the aphasia model achieves the highest F1 on 1), validating it as a prominent feature of aphasia, and the dysarthria model performs the best on 2), which validates it as a prominent feature of dysarthria.

4.2 Semantics Testing

The aim of semantic pressure testing is to evaluate the extent that our models are making predictions based on semantic features, through observing their reaction to altered input descriptions with external or missing information.

- **Semantic Mismatching**: Objects in the patient descriptions are mismatched by both deleting key objects in the cookie theft picture or adding ones that are not in it. The key objects are selected from the TF-IDF ranking of knowledge graph entities extracted in Section 3.1. The performance of both models is a lot more affected by the removal of key objects that have been semantically reinforced by knowledge graphs, than the addition of external objects. This higher sensitivity verifies the effectiveness of knowledge distillation and triple concatenation.
- **Deleting**: The input text is experimented with two different degrees of deleting: 1) only keeping the first sentence, and 2) randomly deleting up to 70% of the text. The aphasia model is more affected by 2) and the dysarthria model is more affected by 1). The results are consistent with our findings in Section 4.1, as the aphasia model's strength with run-on sentences would be negated by random deleting's disruption of sentence structures, while the dysarthria model's strength with fragmented sentences would be irrelevant when there is only one sentence in the input. In addition, the dysarthria model is in general significantly more sensitive to missing texts.
- **Sentence reversal**: We also find that our models could be significantly impacted by reversing the sentence order alone, confirming numerous recent studies that demonstrate BERT-based models' sensitivity to the word or sentence order of the input (Hessel and Schofield, 2021; Pham et al., 2021).

5 Conclusion and Future Work

This work explores the under-researched area of applying NLP to unlabeled patient transcripts for supporting the triage and detection of stroke and stroke mimics. We introduce a storyline-centric approach that leverages data enrichment and knowledge distillation to overcome the lack of big clinical training datasets for automating aphasia and dysarthria detection. Our experiments show that our approach to knowledge distillation has the potential to significantly improve the performance of patient document classification. Nonetheless, we believe that it is possible to further enhance the results in Figure 5 by designing more robust and effective knowledge distillation techniques to integrate transcripts, triples, and graph-theoretic aspects of KGs.

Our ongoing work include: 1) using Sentence Transformers (Reimers and Gurevych, 2019) to further denoise the output of KG extraction, 2) developing solutions to incorporate both semantic knowledge embeddings and graph embeddings in clinical document classification, and 3) recruiting Spanish-speaking patients and healthy volunteers and expanding our storyline-centric pipeline to Spanish language models (Gutierrez-Fandino et al., 2021), to better serve the clinical needs of the Hispanic community in stroke triage and detection.

Limitations

Due to clinical and financial constraints, both the patient and the mTurk sample sizes of our study are still relatively small. This means that we cannot afford to set aside patient data as a hold-out test set, and have to use the validation set for model evaluation. As we work towards enrolling more patients and recruiting more healthy volunteers to

[2] https://www.nidcd.nih.gov/health/aphasia

improve model generalizability, we hope to expand the scope of our pipeline beyond English to serve non-native speaker patients.

One major limitation of the cookie theft picture description task is its lack of equitable assessment for an increasingly diverse patient population. Steinberg et al. (2022) identify gender as a particularly fraught aspect of the picture's expected response, as the rubrics of the initial NIHSS were established from a male-only corpus. Although there is no alternative picture or stroke patient corpus available to our study, we try to ensure the equity of our models by maintaining a gender balance in our patient set, with 136 female patients and 132 male patients. On our patient-only evaluation set, our aphasia model performs significantly better on female patients (AUC=0.909) compared to male patients (AUC=0.702), while the dysarthria model exhibits better performance on male patients (AUC=0.778) than female patients (AUC=0.719). At present, we are unable to draw any definitive conclusions about model equity due to the scale of our data. However, it will be a key area of focus for our future research.

Acknowledgments

We sincerely thank the workshop organizers and anonymous reviewers for their generous time, attention, and feedback. This work was supported by the T. T. & W. F. Chao Foundation and the John S. Dunn Research Foundation.

References

Mark Allen and Dalton Cervo. 2015. *Multi-domain master data management: Advanced MDM and data governance in practice*. Elsevier.

Kristen M. Allison, Yana Yunusov, and Jordan R. Greene. 2019. Shorter sentence length maximizes intelligibility and speech motor performance in persons with dysarthria due to amyotrophic lateral sclerosis." *American journal of speech-language pathology*, 28(1): 96-107.

Randa Ben Ammar and Yassine Ben Ayed. 2018. Speech processing for early Alzheimer's disease diagnosis: Machine learning based approach. In *IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pages 1-8, Aqaba, Jordan.

Tongan Cai, Haomiao Ni, Mingli Yu, Xiaolei Huang, Kelvin Wong, John Volpi, James Z. Wang, and Stephen T. C. Wong. 2022. *DeepStroke*: An efficient stroke screening framework for emergency rooms with multimodal adversarial deep learning. *Medical image analysis*, 80.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2019. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv: 2003.10555*.

Sofia de la Fuente Garcia, Craig Ritchie, and Saturnino Luz. 2020. Artificial intelligence, speech and language processing approaches to monitoring Alzheimer's disease: A systematic review. *Journal of Alzheimer's disease*, 78(4): 1547-1574.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stefan T. Engelter, Michal Gostynski, Susanna Papa, Maya Frei, Claudia Born, Vladeta Ajdacic-Gross, Felix Gutzwiller, and Phillipe A. Lyrer. 2006. Epidemiology of aphasia attributable to first ischemic stroke: Incidence, severity, fluency, etiology, and thrombolysis. *Stroke*, 37(6): 1379-1384.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2021. MarIA: Spanish Language Models. *arXiv preprint arXiv:2107.07253*.

Kristina Lundholm Fors, Kathleen Fraser, and Dimitrios Kokkinakis. 2018. Automated syntactic analysis of language abilities in persons with mild and subjective cognitive impairment. *Studies in health technology and informatics*, 247: 705-709.

Martin Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. https://doi.org/10.5281/zenodo.4461265.

Zhiqiang Guo, Zhenhua Ling, and Yunxia Li. 2019. Detecting Alzheimer's disease from continuous speech using language models. *Jourrnal of Alzheimer's disease*, 70(4): 1163-1174.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*,

pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Jack Hessel and Alexandra Schofield. 2021. How effective is BERT without word ordering? Implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–211, Online. Association for Computational Linguistics.

Howard Kirshner. 2022. Dysarthria and Aphasia of speech. in Bradley and Daroff's *Neurology in Clinical Practice*, 8th Edition, 14: 149-151. United Kingdom, Elsevier.

Anja Kuschmann and Anja Lowit. 2021. Pausing and sentence stress in children with dysarthria due to cerebral palsy. *Folia phoniatrica et logopaedica*, 73(4): 298-307.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Ning Liu, Zhenming Yuan, and Qingfeng Tang. 2021. Improving Alzheimer's disease detection for speech based on feature purification network. *Frontiers in Public Health*, 9.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Brian MacWhinney. 2019. Understanding spoken language through TalkBank. *Behavioral research methods*, 51(4):1919-1927.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning based text classification: A comprehensive review. *ACM computing surveys*, 54(3): 1-40.

Claire Mitchell, Audrey Bowen, Sarah Tyson, Zoe Butterfint, and Paul Conroy. 2017. Interventions for dysarthria due to stroke and other adult-acquired, non-progressive brain injury. The Cochrane database of systematic reviews, 1(1).

Claire Mitchell, Matthew Gittins, Sarah Tyson, Andy Vail, Paul Conroy, Lizz Paley, and Audrey Bowen. 2021. Prevalence of aphasia and dysarthria among inpatient stroke survivors: describing the population, therapy provision and outcomes on discharge. *Aphasiology*, 35(7): 950-960.

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. 2021. Out of Order: How important is the sequential order of words in a sentence in Natural Language Understanding tasks?. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1145–1160, Online. Association for Computational Linguistics.

Yu Qiao, Xuefeng Yin, Daniel Wiechmann, and Elma Kerz. 2021. Alzheimer's disease detection from spontaneous speech through combining linguistic complexity and (dis)fluency features with pretrained language models. *arXiv preprint arXiv:2106.08689*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902– 4912, Online. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Amy Steinberg, Patrick D. Lyden, and Arielle P. Davis. 2022. Bias in Stroke Evaluation: Rethinking the Cookie Theft Picture. Stroke, 53(6): 2123-2125.

P. P. Urban, S. Wicht, G. Vukurevic, C. Fitzek, S. Fitzek, P. Stoeter, C. Massinger, and H. C. Hopf. 2001. Dysarthria in acute ischemic stroke: Lesion topography, clinicoradiologic correlation, and etiology. *Neurology*, 56(8):1

Betty van Aken, Sebastian Herrmann, and Alexander Löser. 2022. What do you see in this patient? Behavioral testing of clinical NLP models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 63-73, Seattle, Washington. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Xintao Wang, Qianyu He, Jiaqing Liang, and Yanghua Xiao. 2022. Language models as knowledge embeddings. *arXiv preprint arXiv:2206.12167*.

Yi Wang, Jiajun Deng, Tianzi Wang, Bo Zheng, Shoukang Hu, Xunying Liu, and Helen Meng. 2023. Exploiting prompt learning with pre-trained language models for Alzheimer's disease detection. *arXiv preprint arXiv:2210.16539*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Kathryn M. Yorkston. 1996. Treatment efficacy: Dysarthria. Journal of speech and hearing research, 39(5): 46-57.