Automated Generation of Storytelling Vocabulary from Photographs for Use in AAC

Mauricio Fontana de Vargas and Karyn Moffatt
School of Information Studies
McGill University, Montreal, Canada
mauricio.fontanadevargas@mail.mcgill.ca, karyn.moffatt@mcgill.ca

Abstract

Research on the application of NLP in symbol-based Augmentative and Alternative Communication (AAC) tools for improving social interaction support is scarce. We contribute a novel method for generating context-related vocabulary from photographs of personally relevant events aimed at supporting people with language impairments in recounting their past experiences. Performance was calculated with information retrieval concepts on the relevance of vocabulary generated for communicating a corpus of 9730 narrative phrases about events depicted in 1946 photographs. In comparison to a baseline generation composed of frequent English words, our method generated vocabulary with a 4.6 gain in mean average precision, regardless of the level of contextual information in the input photographs, and 6.9 for photographs in which contextual information was extracted correctly. We conclude by discussing how our findings provide insights for system optimization and usage.

1 Introduction

Augmentative and Alternative Communication (AAC) tools can enhance communication for non-speaking individuals, thus offering improved social interaction and independence. Well established NLP techniques, such as spell check and word prediction, support those with primarily physical barriers to communication (e.g., adults with ALS) to compose complex and nuanced sentences in orthographic-based systems more efficiently. However, those with developmental disabilities (e.g., autism spectrum disorder, ASD) or lexical and semantic processing impairments that limit their ability to spell out words (e.g., adults with aphasia[1]) must usually rely on less expressive symbol-based systems, for which those techniques offer little support due to unique characteristics of communication with these systems.

Users of symbol-based AAC typically do not construct full, grammatically correct sentences, complete with prepositions and inflections, but rather often only need a few key content words (i.e., nouns, adjectives, verbs)—appearing at any part of the sentence— to supplement other forms of communication, including preserved speech, gestures, or drawings. Such scattered use of vocabulary hinders the typical statistical prediction approach, which relies on patterns learnt from a large training corpus.

Nonetheless, there is much opportunity for improving symbol-based AAC, which is often abandoned because it offers too little communication support relative to the effort required to learn and use (Moffatt et al., 2017).

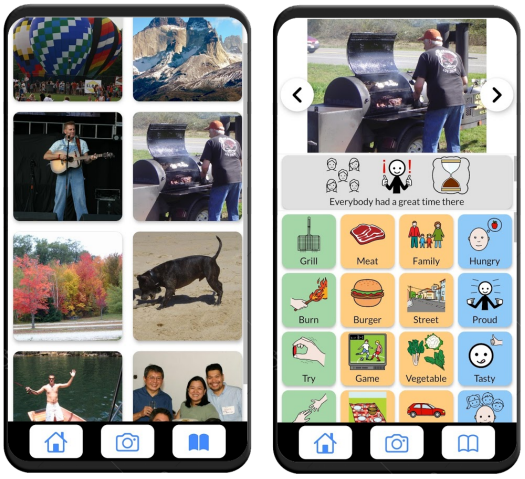


Figure 1: An AAC app design demonstrating how context-related vocabulary generated by our method might be presented for use in subsequent conversations. As in many non-orthographic AACs, vocabulary is represented by images that reproduce computer generated speech when selected; however, unlike the status quo, this design eliminates navigation across complicated hierarchies and the need for pre-programming.

[1] a language disorder mostly often caused by a stroke.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1353–1364
August 1–6, 2021.

Selecting and organizing vocabularies able to attend user's communication needs in a wide variety of contexts and such that they can find words quickly is one of the major challenges (van de Sandt-Koenderman, 2004; Bailey et al., 2006). Alphabetical organizations are not useful, and traditional hierarchical schemes based on abstract categories (e.g., food → apple) are difficult for people with language impairments, making navigation extremely slow for anything but the smallest (least useful) vocabularies. Presenting vocabulary as a flat hierarchy is best (Beukelman et al., 2015; Brock et al., 2017; Wallace and Hux, 2014); however, only a very limited set of options can be displayed, making communication very reliant on having the desired keywords among the available options.

Providing concise situation-relevant vocabularies currently depends on support from a clinician or caregiver to pre-program the device. But such support is often limited or not available, which consequently limits these devices to supporting generic expressions of wants and needs, i.e., functional communication, and not for social interactions involving spontaneous narratives (Waller, 2019).

Generating vocabulary from user's contextual data through Natural Language Generation (NLG) techniques seems an obvious venue to facilitate social interactions. Although NLG has been successfully applied in the context of task-oriented dialogs (He et al., 2017), question answering (Su et al., 2016), text summarization (See et al., 2017), and story generation from photograph sequences (Hsu et al., 2020), it is unclear how these techniques can be adapted to the specific needs of AAC support (Tintarev et al., 2014).

In this paper, we call for more research in the NLP community devoted to language generation for symbol-based AAC systems. We present an overview of the scarce research on the topic and contribute a method that generates vocabulary automatically from a user's photographs to support autobiographical storytelling, demonstrating how it performs under different combination of the system's controllable parameters and a wide range of input photographs.

2 Background and Related Work

2.1 NLP on Orthographic AAC Systems

NLP research on AAC systems has mainly focused on improving the communication rate of orthographic-based tools, primarily via attempts to reduce keystrokes with letter, word, or message prediction, applying n-grams language models on the user input (Swiffin et al., 1985; Garay-Vitoria and Abascal, 2006; Fazly and Hirst, 2003; Trnka et al., 2007; Trnka and McCoy, 2008). Researchers have also explored techniques for improving prediction by including in the language model, some sort of contextual information, such as the topic of conversation (Lesher and Rinkus, 2002; Trnka et al., 2006), the user's location (Garcia et al., 2015), their past utterances (Kristensson et al., 2020; Copestake, 1997; Wandmacher et al., 2008), or their partner's speech (Wisenburn and Higginbotham, 2008). Virtually all commercial text-based high tech AAC devices employ some form of n-gram prediction (Higginbotham et al., 2012).

2.2 The Need for Symbol-based AACs Able to Support Social Interactions

Many people with developmental (e.g., ASD) or acquired disabilities have difficulty using written language, and therefore need support other than orthographic-based AAC. People with expressive aphasia, for example, present lexical and semantic processing impairments that affect their ability to retrieve the names of objects, combine linguistic elements, and use grammar. Nonetheless, they usually have good receptive communication skills and intellectual abilities preserved, and typically desire the ability to communicate complex ideas and share social stories spontaneously, such as describing a recent activity or experience (Garrett, 2005)[2].

To support this population, researchers from the clinical community (McKelvey et al., 2010; Dietz et al., 2006; McKelvey et al., 2007; Beukelman et al., 2015) have successfully explored the presentation of vocabulary associated with personally relevant and highly contextualized photographs, where people, objects, and activities are depicted in their naturally occurring contexts (also known as visual scene displays, VSDs). Evidence indicates greater conversational turn-taking with fewer instances of frustration and navigational errors (Brock et al., 2017), and increased lexical retrieval during activity retell (Mooney et al., 2018), for which participants perceived this kind of support as very helpful.

However, the automation of the language production process to support those social narratives is still highly unexplored. For example, Mooney

[2]We also witnessed this in interactions observed in conversation groups at a local aphasia institute in which the first author participated for 9 months.

et al.'s system CoChat (2018) generates keywords from human input simulating social network comments. NLP was used only to clean the input and identify nouns and frequent words. In consequence, available commercial tools[3] depend on human effort planning and programming relevant vocabulary, leading to lack of spontaneous and independent communication, and requiring a great amount of time from caregivers (Drager et al., 2019).

2.3 NLG for AAC Systems

Generating language for AAC systems is highly different from typical NLG usage, mainly because the goal of AAC is to provide support for communicating users thoughts, and not to replace the user by an automatic communicator (Tintarev et al., 2014).

The Compansion system (Demasco and McCoy, 1992; McCoy et al., 1998), was one of the first attempts to apply NLG towards that goal. It was designed to produce grammatically correct sentences from incomplete user input using a small domain model. Although Compansion was dedicated to functional communication, its concept of using a domain knowledge served as a stepping stone to Dempster et al.'s system aimed at generating conversational utterances (2010). In their prototype, users populated a personal knowledge base by recording where, when, and with whom they performed an activity shortly after its end. Through a template-driven system, users' knowledge was converted into conversational utterances organized on topics that could be accessed during subsequent conversations. This work showed promising results on how NLG can be able to support social dialogues and increase participation of AAC users. However, their system still required considerable manual linguistic input from users.

Automatic generation of storytelling vocabulary has been successfully explored by researchers (Reiter, 2007; Black et al., 2010; Tintarev et al., 2016) to support children with limited memory or with physical and intellectual impairment telling "how was school today" to their parents. In their project, raw sensor data from passive RFID tags relating to locations, objects, and people was aggregated into events, and then transformed to coherent personal narratives using a domain knowledge containing the school timetable and the RFID tags mapping.

To provide just-in-time vocabularies that attend to emergent needs and are not tied to a specific scenario (e.g., school), Demmans Epp et al. (2012) explored the use of information retrieval algorithms on internet-accessible corpora such as websites, dictionaries, and Wikipedia pages related to the user's current location or conversation topic. Although this approach was useful for augmenting a base vocabulary with context-specific terms, it is limited to locations (e.g., retail locations) for which internet-accessible corpora are likely to exist.

3 Vocabulary Generation Method

Our method generates a rank of key words and short narrative phrases from a single[4] input photo for scaffolding storytelling. It was designed to be used as the back end of interactive AAC systems in which relevant vocabulary is associated with a main photograph, such as Mooney et al.'s CoChat, or as in the example design shown in Fig. 1.

We used VIST-TRAIN, a sub-set of the visual storytelling dataset VIST (Huang et al., 2016) as the main source for vocabulary generation. VIST-TRAIN encompasses 80% of the entire dataset, and is composed of 65,394 photos of personal events, grouped in 16,168 stories. Each photo is annotated with descriptions and narrative phrases that are part of a story, created by Amazon Mechanical Turk workers. We judged VIST to be a good source of vocabulary because i) photos were extracted from personal Flickr albums on a wide range of "storyable" events, related to 69 topics (e.g., graduation, building a house), ii) associated vocabulary is representative of storytelling and, iii) stories and photo descriptions were constructed by a large number (1907) of workers under a rigorous procedure.

The generation process is composed of five steps, as detailed below and illustrated in Fig 2. We explore different implementations for some of the steps, represented by the system's controllable parameters emphasized with bold italic formatting throughout the paper. The different combination of those parameters are evaluated in the next section.

3.1 Scene Understanding

The first step extracts contextual information from the photograph in the form of a high-level, human-like description of the scene (i.e., caption) using the computer vision technique from Fang et al. (2015). Captioning was chosen over pure object detection and labelling due to the necessity of communicat-

[3] e.g., Tobii Dynavox Snap Scene

[4] to reduce the requirements on users, who may feel discouraged if multiple photos of the event are needed

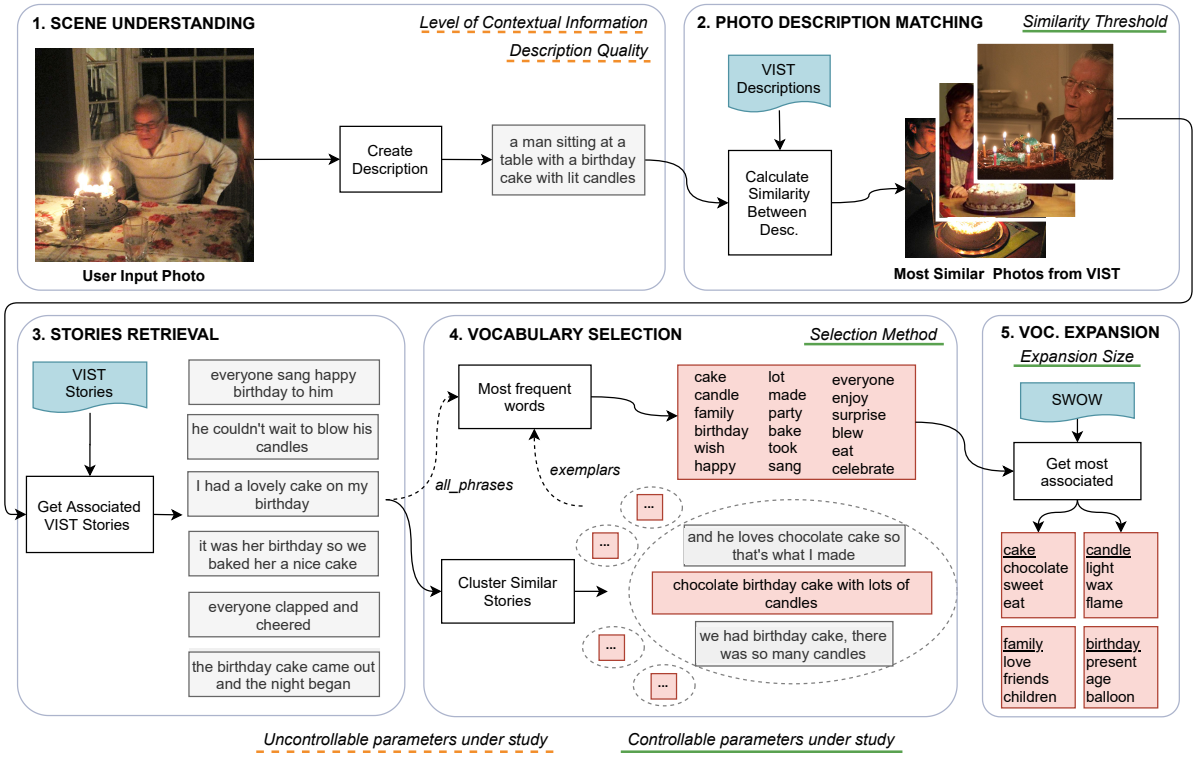


Figure 2: Our method. Words and phrases highlighted in red are generated from the input photograph.

ing more abstract concepts such as the actions being performed and the interactions between the objects, people, and environment during storytelling.

3.2 Photo Description Matching

This step finds the subset of VIST-TRAIN photos most similar to the user input by calculating the sentence similarity between the input photo description and all VIST-TRAIN photos descriptions. All photos with description similarity higher than the parameter ***Similarity Threshold*** are selected for processing in the next step, with an upper limit of 30 photos. Sentence similarity is defined as the soft cosine similarity (Sidorov et al., 2014)[5] on a bag-of-words representation of the sentences using Word2Vec embeddings, after removing stop words[6]. Soft cosine was chosen as similarity measure due to its ability to capture the semantic relatedness between different words. This strategy was motivated by the fact that soft cosine similarity with Word2Vec was effective for finding similar sentences on question-answering systems, achieving the best performance at the SemEval-2017 Task 3 (Charlet and Damnati, 2017). Similarity based on entire documents (e.g., Doc2Vec) was not used because it would require a much larger (at present, nonexistent) training corpus to create proper document embeddings, and there are no pre-trained sentence embeddings trained exclusively on photo descriptions.

3.3 Stories Retrieval

All narrative sentences associated with the selected photos are retrieved for processing in the next stage. The number of sentences per photo varies from 1 to 5 ($\mu = 3.1, \sigma = 1.4$).

3.4 Vocabulary Selection

This step identifies a group of representative sentences and words from the retrieved set by applying the Affinity Propagation[7] clustering (Frey and Dueck, 2007)—able to generate clusters with less error than other exemplar-based algorithms and not requiring a predetermined the number of clusters. The final set of generated phrases is formed by these cluster's exemplars, ranked according to their respective clusters size. By definition, this strategy results in phrases covering the wide range of semantics present in the set of retrieved phrases, while at the same time removing redundant (i.e., very similar) phrases. In case of

[5]Gensim library implementation

[6]as defined by the Natural Language Toolkit (NLTK)

[7]damping: 0.5, max. iter: 200, convergence iter.: 15

non-convergence (< 3% in our evaluation), the set of recommended phrases is formed by ranking all phrases according to the sum of their soft cosine similarity against all other phrases retrieved. The generated base vocabulary is formed by a rank of the word frequencies after filtering-out stop words and applying a porter stemmer to merge different variations (e.g., worked, working → work). The parameter ***Selection Method*** determines whether frequencies are calculated considering all retrieved phrases (ALL_PHRASES) or only clusters' exemplars (EXEMPLARS).

3.5 Vocabulary Expansion

The goal of this step is to diversify the base vocabulary derived from VIST-TRAIN to increase communication flexibility. Thus, to find words that are related to, but distinct from the initial concept (e.g., cake → sweet), our method uses a model of the human mental lexicon as a secondary source of vocabulary. In this model, SWOW (De Deyne et al., 2019), words are connected with a certain strength representing their relatedness constructed from data of word-association experiments of over 90,000 participants. Therefore, unlike embeddings, SWOW encodes mental representations free from the basic demands of communication.

This strategy was motivated by the fact that word association data was successfully applied in a controlled study to support people with aphasia navigating related words more effectively (Nikolova et al., 2010), and that evidence from cognitive science research indicates that the network formed by associations in SWOW presents a widespread thematic structure, rather than taxonomic, with words strongly associated often occurring in the same situation (e.g., pick-strawberry; candle-church) (De Deyne et al., 2015) . This last step expands the initial set of base vocabulary by adding, for each word, the most strongly associated words in SWOW data. The system parameter ***Expansion Size*** determines how many words from SWOW are added for each word in the base vocabulary set. Repeated words are not included.

4 Evaluation Experiment

The goal of our evaluation is to understand how our design choices, represented by the system *controllable parameters*, along with uncontrollable factors related to the input photograph (i.e., *uncontrollable parameters*), affect the system's performance. Thus, we compared the relevance of vocabulary generated under different combinations of these parameters to investigate the following specific research questions:

1. What combination of controllable system parameters related to the base vocabulary generation optimizes performance?
2. How does the level of contextual information in the input photo affect performance?
3. How does the quality of the contextual description inferred from the input photo affect performance?
4. How does the level of contextual information in the input photo affect the quality of the inferred description?
5. What is the effect of expanding the base generated vocabulary with words from a mental lexicon model on the system's performance?

4.1 Performance Metrics

Considering the AAC application usage scenario, the performance of vocabulary generation can be conceptualized by the combination of two factors: i) communication flexibility, i.e., whether vocabulary needed for composing messages about a specific experience is provided, and ii) communication ease, i.e., the difficulty in finding a particular word among all options generated. These two factors directly map to the information retrieval concepts of precision (P) and recall (R) as a perfect algorithm would provide all words the user needs to communicate the desired message (R = 1), and would not contain any irrelevant vocabulary (P = 1), thereby minimizing the need for scanning. In contrast, the worst algorithm would provide only irrelevant vocabulary ($P = R = 0$).

Therefore, we tackle the vocabulary generation evaluation as an information retrieval problem, where the input photo is treated as the user query, generated words and phrases are treated as retrieved documents, and crowd sourced narrative sentences about the photograph are the relevant documents, i.e., ground truth (as detailed in Section 4.2). For each input photo, difficulty in finding vocabulary and communication flexibility are operationalized as P and R, respectively:

$$P(n) = \frac{|\{rel\_words\} \cap \{G_n\}|}{n}$$

$$R(n) = \frac{|\{rel\_words\} \cap \{G_n\}|}{|\{rel\_words\}|}$$

where n is the number of words displayed to the user, rel_words are the words in the groundtruth sentences, and G_n are the top n words in the generated vocabulary rank. We also calculated the F_1, a common information retrieval measure that captures the trade-off between P and R:

$$F_1(n) = 2 \times \frac{P(n) \times R(n)}{P(n) + R(n)}$$

We calculated these metrics for all $n \in [1, 100]$, and constructed the P-R curves with the arithmetic mean values of P, R, and F_1 across all input photographs under analysis. In contrast to BLEU/METEOR metrics, this analysis allows us to clearly demonstrate trade-offs between the difficulty finding a word among options and communication flexibility, which is important because the number of displayed items will vary for each user.

To obtain a single measure of system performance across this entire interval, considering all input photos, we approximate the area under the P-R curves by calculating the mean average precision:

$$mAP = \sum_{n=1}^{100} P(n)(R(n) - R(n-1))$$

4.2 Data

As input photographs and groundtruth sentences, we used VIST-VAL, a sub-set of VIST not employed in our method that contains 8034 photos aligned with crowd sourced stories. We selected all photos from VIST-VAL containing the maximum number of sentences available (5) to act as our input photographs, resulting in 1946 photos. The ground-truth vocabulary for each photograph was formed by joining the five associated narrative phrases (9730 in total), after removing stop words.

4.3 Specific Procedures

Controllable Parameters - Base Vocab. (RQ1). We defined four configurations of parameters by crossing two extreme values of ***Similarity Threshold***, i.e., *0* and *best* (highest similarity score among all VIST-VAL) with the ***Selection Method*** *all_phrases* and *exemplars*, resulting in four configurations: 0_ALL, 0_EXEMPLARS, BEST_ALL, BEST_EXEMPLARS. ***Expansion size*** was set to 0 in all configurations. In the absence of similar AAC generation systems to compare our method to, we created a BASELINE generation formed by a rank of the most frequent words from the Corpus of Contemporary American English (COCA) (Davies, 2009) without stop words. We adopted this baseline because current AAC tools are commonly built on word usage frequency data (Renvall et al., 2013).

The optimal values for the parameters established in this analysis were applied in subsequent analyses.

Contextual Information Level (RQ2, RQ4). To investigate the variability caused by different input photographs, we adopted the concept of context richness from Beukelman et al. (2015). The first author scored each photo from 0–3 based on the number of contextual categories (environment, people/object, activity) it clearly depicts (0 when ambiguous). To validate these annotations, someone unfamiliar with the study also scored a subset of 514 photos (27.8% of the dataset)[8]. Krippendorff's alpha reliability score was 0.82, indicating strong agreement between raters (Krippendorff, 2004).

Context Description Quality (RQ3, RQ4). The first author scored each photo description from 0 to 3 as follows: 0) not generated or completely unrelated; 1) misses most important elements OR contains most of important elements and a few unrelated elements; 2) contains most of important elements OR all important elements and a few unrelated elements; 3) contains all important elements in the photo and does not contain any unrelated elements. As for contextual information level, a subset of 514 were scored by someone unfamiliar with the study. Krippendorff's alpha reliability score was 0.88, confirming strong agreement.

Effect of Vocabulary Expansion (RQ5). We created 24 pairs of configurations by combining different base vocabulary sizes (5, 10, 15, 20, 25, 30) with the expansion sizes (0, 1, 2, 3). The configuration [5-2], for example, contains five base words plus two expanded words per base word, resulting in a maximum of 15 words (or less if expanded words were already in the base set).

4.4 Results

RQ1. To better illustrate the differences in performance, Fig. 3 presents the P-R curves, while Table 1 shows the mAP and maximum P and R mean values for the pairs of parameters values under investigation, in comparison to the baseline. Overall, 0_ALL results in the best performance,

[8] all annotations are available at https://doi.org/10.5683/SP2/NVI701

with an mAP 4.6 times greater than the baseline, and 1.8 greater than the the worst configuration, BEST_EXEMPLARS.

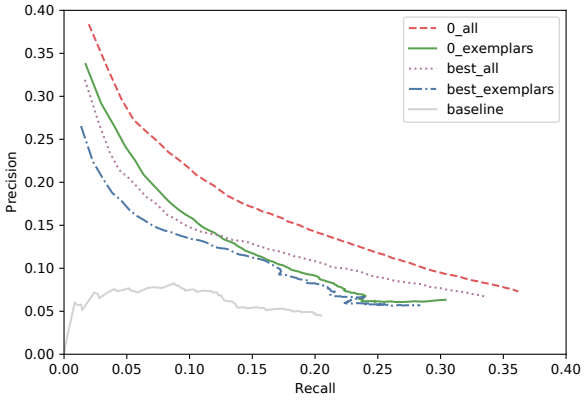


Figure 3: P-R curves for different configurations of system's parameters, calculated for all $n \in [1, 100]$.

Configuration	mAP	mAP gain	max P	max R
0_ALL	.058	4.61	.38	.36
0_EXEMP	.039	3.10	.34	.30
BEST_ALL	.042	3.35	.32	.33
BEST_EXEMP	.032	2.52	.27	.28
BASELINE	.013	1.00	.08	.20

Table 1: Performance under different configurations.

RQ2. In our input dataset, the proportion of photos according to their context richness score was: 8%(0), 54%(1), 30%(2), 8%(3). A Mann-Whitney U test indicated a significant difference on P and R only between photos with context richness 0 and the remaining levels ($p < .002$). Table 2 shows the mean performance metrics according to level of contextual information.

Context Level	mAP	mAP gain	max P	max R
3	.056	4.44	.43	.37
2	.060	4.72	.38	.36
1	.058	4.57	.38	.36
0	.045	3.54	.29	.23
BASELINE	.013	1.00	.08	.20

Table 2: Mean performance according to the level of contextual information in the input photos.

RQ3. The distribution of input photos across context description quality scores was: 16%(0), 16%(1), 30%(2), 38%(3). We plot the P-R curves according to the context description quality scores in Fig. 4, and summarize performance metrics in Table 3. A Mann-Whitney U test indicated no significant differences between photo quality 1 and 2 ($p > .2$). However, photos with description quality 3 significantly outperformed the other groups ($p < .001$), and quality 0 photos performed significantly worse than all other groups ($p < .001$).

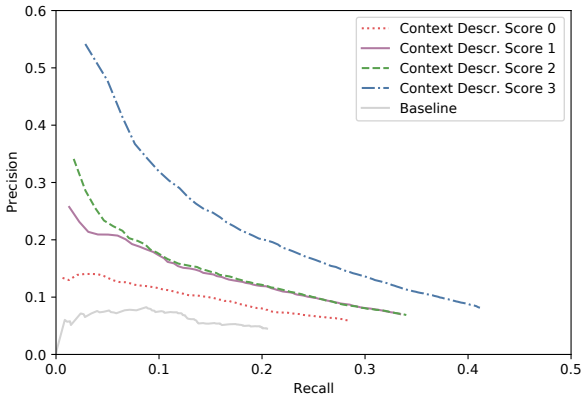
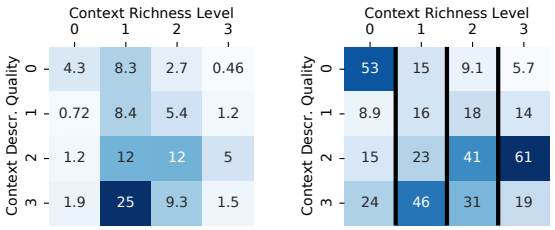


Figure 4: Precision-recall curves according to context description quality, under the configuration 0_ALL.

Descr. Quality	mAP	mAP gain	max P	max R
3	.086	6.86	.54	.41
2	.048	3.77	.34	.34
1	.045	3.57	.26	.33
0	.028	2.21	.14	.29
BASELINE	.013	1.00	.08	.20

Table 3: Mean performance metrics according to the input photos' description quality.

RQ4. Fig. 5 illustrates the relationship between the level of contextual information in the input photos and the quality of the photos descriptions generated using machine-learning.



(a) Percentages relative to all photos (1946)

(b) Percentages relative to photos with same context richness level

Figure 5: Distribution of input photos by contextual richness level and generated description quality

As expected, photos with ambiguous contextual information (level= 0) most often received bad captions (53%). As context richness increased, the relative proportion of photos with good descriptions (scores 2 or 3) also increased (39%, 69%, 72%,

80%), but the relative proportion of perfect descriptions (quality = 3) decreased (46%, 31%, 19%). Photos depicting only one type of contextual information (location, person/object, activity) resulted in the best descriptions: 46% received perfect descriptions, and 66% of all perfect descriptions were given to them. However, when compared to photos with more contextual information, they presented the highest relative proportion of very bad captions (15% vs 9.1% and 5.7%).

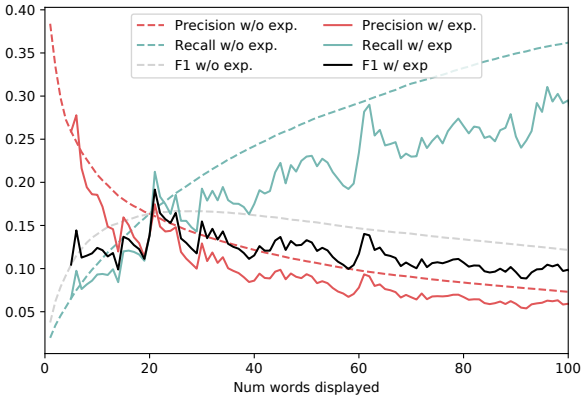


Figure 6: Comparison between generation with and without vocabulary expansion.

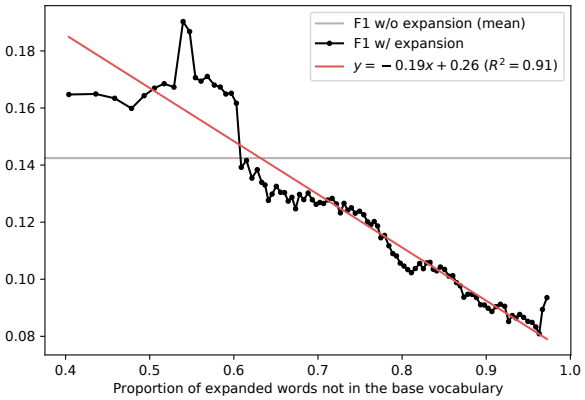


Figure 7: Impact of the intersection between base and expanded vocabulary on performance.

RQ5. Fig. 6 compares the performance of different combinations of base vocabulary and expansion sizes against base vocabulary only, in function of the number of words displayed n. In general, for a given n, generation without expansion resulted in superior performance. However, on configurations for which a high proportion of expanded words were already in the base vocabulary (e.g., $n = 6$, 21, 61), expansion presented similar or even better F_1 scores than the base vocabulary on its own.

To better understand this phenomenon, we plot the F_1 score, averaged across all photos, in function of the proportion of expansion words not present in the base vocabulary during generation (Fig. 7). The mean F_1 for generation without word expansion is also plotted for comparison.

We found that word expansion is able to bring improvement in performance when less than 60% of the expansion words are included in the final generated vocabulary, or in other words, when more than 40% of expansion words is already in the base vocabulary. The tendency is that, the lower the proportion of expansion words not in the base vocabulary, the higher the performance.

5 Discussion

The design space for generating AAC storytelling vocabulary directly from photographs is vast and under explored. Design decisions for individual system components will impact other components and ultimately the overall system effectiveness, and therefore cannot be arbitrary. Without a rigorous performance evaluation on different configurations of parameters, users would be at risk of using a flawed or under optimized system, which could lead to user frustration and abandonment, and cause confounds that obscure whether failures are due to the need for algorithmic tuning or mismatch between the intended support and user needs.

The study of controllable parameters (RQ1, 5) demonstrated that **our method is able to provide relevant vocabulary**, and showed how it can be used to optimize the system and identify areas for further improvement. The exploration of uncontrollable parameters (RQ2, 3, 4) helped illustrate the likely variation in system performance during real world usage (i.e., wide variety of input photos), allowing us to better anticipate potential problems or pitfalls and understand requirements for use.

The similar performance across photos with different levels of contextual information (RQ2) suggests that **our method is robust to variations in the input photograph**. Users will not need to be instructed to take photographs following specific requirements, e.g., "photos should demonstrate an action" or "photos should depict objects only". The similar levels of performance is explained by the pattern observed in the RQ4 analysis; the more elements a photo contains, the better knowledge the machine learning has to infer the central aspect of the photo, but at the same time, the harder it is to capture each and every element. In addition, an

element wrongly identified will have less impact on the overall scene understanding since other elements complement the description. An example would be a photo of a birthday party, in which the machine-learning platform is able to infer the central concept (birthday) from the several elements depicted (e.g., cake, candles, balloons), but misses some of the details (e.g. drinks). On the other hand, simplistic photos will rarely lead to elements being cut out, but the computer vision technique will have more variability when performing the inferences, leading to erroneous descriptions more often.

On the other hand, the quality of generated vocabulary was strongly dependent on the computer vision technique employed to extract contextual information about the scene (RQ3) . When a wrong description is generated, the subsequent steps of the algorithm are misled and therefore generate vocabulary less relevant for retelling the scene depicted in the photograph. Nonetheless, even in this case, an AAC device using our method would provide vocabulary more relevant than if the most frequent English words were provided. Since photos for which the computer vision technique was able to correctly identify all contextual elements resulted in substantial performance gain, we encourage further exploration of this component. An option would be to use a higher number of raw context labels instead of the single human-like description employed in this work.

Our vocabulary expansion analysis (RQ5) provide valuable insights into how the combination of multiple lexicon sources can generate more relevant vocabulary. **The most promising approach was to combine the visual-to-story dataset with strongly associated words from a mental-lexicon model, but only when there was high intersection between the two vocabularies.**

5.1 Limitations and Future Work

Although VIST contains a very large range of events, one limitation is that it is unlikely to cover all possible scenarios, and may not accurately reflect AAC communication. However, in the absence of an appropriate AAC-specific corpora (a known issue in the community), we believe the VIST dataset can meaningfully represent the vocabulary needed for scaffolding storytelling. In addition, we do not expect the performance gains observed will directly translate to the same gains in usability. Our goal was to understand fundamental questions necessary for advancing to a usability study, helping fine-tune system components before introducing them to users, avoiding unnecessary interactions with identifiably poor designs. Our approach also enables larger numbers of parameters to be examined. The low level of social participation commonly observed among people with aphasia, combined with the rate-limited nature of AAC, would require field experiments lasting an impractical amount of time to produce sufficient data to comprehensively explore possible combinations of parameters (Kristensson et al., 2020).

As a potential improvement to our method, Sent2Vec trained with BERT may better represent sentence structure and words context for finding similar photo descriptions in step 2 than our use of soft cosine with Word2Vec. Another option would be the use of query expansion to enrich the descriptions. We encourage the exploration of the vast array of strategies for tackling the vocabulary generation process for AAC.

6 Conclusion

Developing a photo-to-story vocabulary AAC system presents two challenges; a NLP one in how to generate such vocabularies, and a Human-Computer-Interaction (HCI) one in how to use such vocabulary to offer interactive language support. In this work, we tackle the first challenge.

We demonstrated that our method is able to generate vocabulary with reasonable levels of recall and precision, regardless of the level of contextual information in the input photograph, illustrated the likely variation in system performance during real world usage, and provided meaningful insights for fine tuning the algorithm, enabling us to move to the next phase of designing and evaluating, with AAC users, our mobile interactive application.

Acknowledgments

This research was funded by the Fonds de Recherche du Québec - Nature et Technologies (FRQNT), the Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN-2018-06130], the Canada Research Chairs Program (CRC), and by AGE-WELL NCE, Canada's technology and aging network.

References

Rita L Bailey, Howard P ParetteJr, Julia B Stoner, Maureen E Angell, and Kathleen Carroll. 2006. Family members' perceptions of augmentative and alternative communication device use. *Language, Speech, and Hearing Services in Schools*, 37(1).

David R Beukelman, Karen Hux, Aimee Dietz, Miechelle McKelvey, and Kristy Weissling. 2015. Using visual scene displays as communication support options for people with chronic, severe aphasia: A summary of AAC research and future research directions. *Augmentative and Alternative Communication*, 31(3):234–245.

Rolf Black, Joseph Reddington, Ehud Reiter, Nava Tintarev, and Annalu Waller. 2010. Using NLG and sensors to support personal narrative for children with complex communication needs. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, pages 1–9.

Kris Brock, Rajinder Koul, Melinda Corwin, and Ralf Schlosser. 2017. A comparison of visual scene and grid displays for people with chronic aphasia: A pilot study to improve communication using aac. *Aphasiology*, 31(11):1282–1306.

Delphine Charlet and Geraldine Damnati. 2017. Simbow at semeval-2017 task 3: Soft-cosine semantic similarity between questions for community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 315–319.

Ann Copestake. 1997. Augmented and alternative NLP techniques for augmentative and alternative communication. In *Natural Language Processing for Communication Aids*.

Mark Davies. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190.

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, 51(3):987–1006.

Simon De Deyne, Steven Verheyen, Amy Perfors, and Daniel J Navarro. 2015. Evidence for widespread thematic structure in the mental lexicon. In *CogSci*.

Patrick W Demasco and Kathleen F McCoy. 1992. Generating text from compressed input: An intelligent interface for people with severe motor impairments. *Communications of the ACM*, 35(5):68–78.

Carrie Demmans Epp, Justin Djordjevic, Shimu Wu, Karyn Moffatt, and Ronald M Baecker. 2012. Towards providing just-in-time vocabulary support for assistive and augmentative communication. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 33–36.

Martin Dempster, Norman Alm, and Ehud Reiter. 2010. Automatic generation of conversational utterances and narrative for augmentative and alternative communication: A prototype system. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, pages 10–18.

Aimee Dietz, Miechelle McKelvey, and David R Beukelman. 2006. Visual scene displays (VSD): New AAC interfaces for persons with aphasia. *Perspectives on Augmentative and Alternative Communication*, 15(1):13–17.

Kathryn DR Drager, Janice Light, Jessica Currall, Nimisha Muttiah, Vanessa Smith, Danielle Kreis, Alyssa Nilam-Hall, Daniel Parratt, Kaitlin Schuessler, Kaitlin Shermetta, et al. 2019. AAC technologies with visual scene displays and "just in time" programming and symbolic communication turns expressed by students with severe disability. *Journal of intellectual & developmental disability*, 44(3):321–336.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482.

Afsaneh Fazly and Graeme Hirst. 2003. Testing the efficacy of part-of-speech information in word completion. In *Proceedings of the 2003 EACL Workshop on Language Modeling for Text Entry Methods*.

Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science*, 315(5814):972–976.

Nestor Garay-Vitoria and Julio Abascal. 2006. Text prediction systems: A survey. *Universal Access in the Information Society*, 4(3):188–203.

Luís Filipe Garcia, Luís Caldas De Oliveira, and David Martins De Matos. 2015. Measuring the performance of a location-aware text prediction system. *ACM Transactions on Accessible Computing (TACCESS)*, 7(1):1–29.

Kathryn L Garrett. 2005. Adults with severe aphasia. In David R Beukelman and Pat Mirenda, editors, *Augmentative and alternative communication for children and adults with complex communication needs*, pages 467–504. Paul H. Brookes, Baltimore.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. *arXiv preprint arXiv:1704.07130*.

D Jeffery Higginbotham, Gregory W Lesher, Bryan J Moulton, and Brian Roark. 2012. The application of natural language processing to augmentative and alternative communication. *Assistive Technology*, 24(1):14–24.

Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao Huang, and Lun-Wei Ku. 2020. Knowledge-enriched visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7952–7960.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Per Ola Kristensson, James Lilley, Rolf Black, and Annalu Waller. 2020. A design engineering approach for quantitatively exploring context-aware sentence retrieval for nonspeaking individuals with motor disabilities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–11.

Gregory W Lesher and Gerard J Rinkus. 2002. Domain-specific word prediction for augmentative communication. In *Proceedings of the RESNA 2002 Annual Conference*.

Kathleen F McCoy, Christopher A Pennington, and Arlene Luberoff Badman. 1998. Compansion: From research prototype to practical integration. *Natural Language Engineering*, 4(1):73–95.

Miechelle L McKelvey, Aimee R Dietz, Karen Hux, Kristy Weissling, and David R Beukelman. 2007. Performance of a person with chronic aphasia using personal and contextual pictures in a visual scene display prototype. *Journal of Medical Speech Language Pathology*, 15(3):305.

Miechelle L McKelvey, Karen Hux, Aimee Dietz, and David R Beukelman. 2010. Impact of personal relevance and contextualization on word-picture matching by people with aphasia. *American Journal of Speech-Language Pathology*.

Karyn Moffatt, Golnoosh Pourshahid, and Ronald M Baecker. 2017. Augmentative and alternative communication devices for aphasia: The emerging role of "smart" mobile devices. *Universal Access in the Information Society*, 16(1):115–128.

Aimee Mooney, Steven Bedrick, Glory Noethe, Scott Spaulding, and Melanie Fried-Oken. 2018. Mobile technology to support lexical retrieval during activity retell in primary progressive aphasia. *Aphasiology*, 32(6):666–692.

Sonya Nikolova, Marilyn Tremaine, and Perry R Cook. 2010. Click on bake to get cookies: Guiding word-finding with semantic associations. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*, pages 155–162.

Ehud Reiter. 2007. An architecture for data-to-text systems. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 97–104.

Kati Renvall, Lyndsey Nickels, and Bronwyn Davidson. 2013. Functionally relevant items in the treatment of aphasia (part ii): Further perspectives and specific tools. *Aphasiology*, 27(6):651–677.

Mieke van de Sandt-Koenderman. 2004. High-tech aac and aphasia: Widening horizons? *Aphasiology*, 18(3):245–263.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. 2014. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3):491–504.

Yu Su, Huan Sun, Brian Sadler, Mudhakar Srivatsa, Izzeddin Gür, Zenghui Yan, and Xifeng Yan. 2016. On generating characteristic-rich question sets for qa evaluation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 562–572.

Andrew L Swiffin, J Adrian Pickering, John L Arnott, and Alan F Newell. 1985. PAL: An effort efficient portable communication aid and keyboard emulator. In *8th Annual Conference on Rehabilitation Technology, Technology-A Bridge to Independence. RESNA'85. Memphis, Tennessee*, pages 197–199. Rehabilitation Engineering Society of North America.

Nava Tintarev, Ehud Reiter, Rolf Black, and Annalu Waller. 2014. Natural language generation for augmentative and assistive technologies. In *Natural Language Generation in Interactive Systems*, pages 252–277. Cambridge University Press.

Nava Tintarev, Ehud Reiter, Rolf Black, Annalu Waller, and Joe Reddington. 2016. Personal storytelling: Using natural language generation for children with complex communication needs, in the wild.... *International Journal of Human-Computer Studies*, 92:1–16.

Keith Trnka and Kathleen F McCoy. 2008. Evaluating word prediction: Framing keystroke savings. In *Proceedings of ACL-08: HLT, Short Papers*, pages 261–264.

Keith Trnka, Debra Yarrington, John McCaw, Kathleen F McCoy, and Christopher Pennington. 2007. The effects of word prediction on communication rate for AAC. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 173–176.

Keith Trnka, Debra Yarrington, Kathleen McCoy, and Christopher Pennington. 2006. Topic modeling in fringe word prediction for AAC. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 276–278.

Sarah E Wallace and Karen Hux. 2014. Effect of two layouts on high technology AAC navigation and content location by people with aphasia. *Disability and Rehabilitation: Assistive Technology*, 9(2):173–182.

Annalu Waller. 2019. Telling tales: Unlocking the potential of AAC technologies. *International journal of language & communication disorders*, 54(2):159–169.

Tonio Wandmacher, Jean-Yves Antoine, Franck Poirier, and Jean-Paul Départe. 2008. Sibylle, an assistive communication system adapting to the context and its user. *ACM Transactions on Accessible Computing (TACCESS)*, 1(1):1–30.

Bruce Wisenburn and D Jeffery Higginbotham. 2008. An AAC application using speaking partner speech recognition to automatically produce contextually relevant utterances: Objective results. *Augmentative and alternative communication*, 24(2):100–109.