Gaze-based Multimodal Meaning Recovery for Noisy / Complex Environments

Özge Alaçam
oezge.alacam@uni-hamburg.de
Language Technology, Universität Hamburg
Hamburg, Germany

Ganeshan Malhotra*
BITS Pilani Goa
Pilani, India

Eugen Ruppert
Base.camp, Universität Hamburg
Hamburg, Germany

Chris Biemann
Language Technology, Universität Hamburg
Hamburg, Germany

ABSTRACT

Reference resolution is an important problem that has enormous practical implications in daily life, for example in recovering the intended meaning in communication when the environment is noisy (acoustic noise in the spoken channel, or clutter / occlusion in the visual world). Recent literature indicates that cross-modal processing of all the contributive modalities improves the reference resolution in such settings. In this paper, we investigate the contribution of the eye-tracking methodology, a substantial but underrepresented component of face-to-face communication in NLP systems, to recover the meaning in noisy settings. We integrate gaze features into state-of-the-art language models and test the model on data where parts of the sentences are masked, mimicking noise in the acoustic channel. The results indicate that eye movements can compensate for the missing information in the situation and support communication when language and visual modality fail.

KEYWORDS

meaning recovery; gaze detection; multimodal communication

ACM Reference Format:
Özge Alaçam, Ganeshan Malhotra, Eugen Ruppert, and Chris Biemann. 2021. Gaze-based Multimodal Meaning Recovery for Noisy / Complex Environments. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3462244.3481002

1 INTRODUCTION

Considering the drastic increase in the use of assistive technologies such as smart speakers or, in a more advanced way, collaborative robots that can engage in communication, there is a considerable need for NLP (Natural Language Processing) models that can process and comprehend a wide range of situations. To achieve this, multimodal integration of all available modalities in the communication including eye-movements plays a crucial role.

In a task-oriented setting (e.g. helper robots completing a given task), one of the crucial NLP processes is to extract the intention of the speaker. Figure 1 illustrates a simple case, where the user's command *"Please bring me the glasses!"* conveys a lexical ambiguity (the word *glasses* might refer to either the drink glasses on the far side table or the eye-glasses located on the middle side table). A more realistic scenario might even contain several objects of the same types with various properties. Some of them could even be (partially) occluded from the viewer's perspective. The environments usually also include several people and their interactions with the objects. In such cases, incorporating contextual information plays a crucial role in determining the objects to be referred to and in accomplishing the task (see [1] for a review). Revisiting Figure 1, there is a newspaper in front of the user, so the possible action could be reading. This cue might increase the probability of eye-glasses as the intended object. Moreover, there is already a drink glass next to the user, and this cue might decrease the probability of choosing drink glasses. However, incorporating such knowledge requires high-quality extraction of object labels and relations in the scene, of which state-of-the-art (SOTA) computer vision algorithms still fall short despite recent impressive gains in performance. This situation-specific information further needs to be integrated into the language model (LM) via some reasoning components to be able to perform the above-mentioned reasoning. In addition to the problems that might occur during information extraction, another possibility that can hinder successful communication is noise in

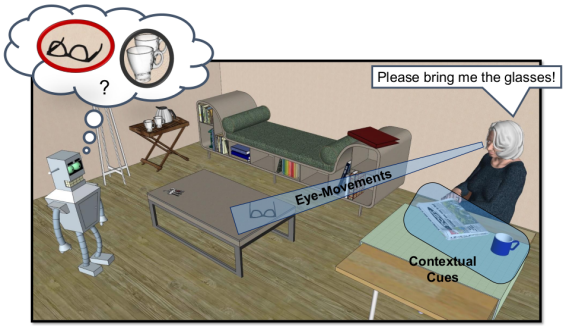


Figure 1: An example for a helper robot scenario.

*Remote research-intern at Language Technology Group, University of Hamburg

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3481002

the environment. Noise in communication can originate from various sources. It can be linguistic noise (e.g. complex attachments, fragmented sentences), visual ambiguities (e.g. clutter in the environment, occlusions) or an acoustic noise (speech recognition errors). Being able to follow the gaze of interlocutors is one of the advantages of human–human multimodal communication, a feature that human–robot interaction can benefit immensely from, particularly when the language or visual modules fail.

The human language processing system integrates information from various modalities. Mainstream NLP systems – however – still usually employ uni-modal approaches, where the performance is highly dependent on the completeness of the language modality. However, spontaneously spoken utterances typically lack completeness. Especially in noisy settings, where the informativeness of the language modality is low, other informative cues coming from the environment and from the communicational partners, such as gaze direction, or deictic/symbolic gestures, provide more reliable information. Most recently, not just high-end assistive technologies but also daily-use devices like phones or laptops begin to utilize eye-tracking technology [7, 15, 28]. Therefore, incorporating eye-movements in our language comprehension models is inevitable for NLP emerging from these developments, and this motivates systematic research on the interaction of different modalities.

2 MULTIMODAL INTERACTIONS FOR MEANING RECOVERY

Due to long-range dependencies, flexible word order, syntactic incompleteness and fragmentation of spoken language [e.g. 14], extraction of speaker intention requires the integration of various information sources, thus it is a highly complex process. These sources can be both low-level perceptual channels like sounds and visual features, or top-down sources like commonsense knowledge, concepts and relations, and even some other multimodal cues like eye-movements, gestures and haptic movements.

The issue of how to comprehend noisy linguistic input and reconstruct the intended meaning has been a focus of both psycholinguistic and computational line of research, [e.g. 19]. According to the noisy-channel model [11], the sentence comprehension mechanism integrates all the information at the syntactic, semantic and discourse level from the existing words and uses this linguistic evidence to predict the missing parts and infer the possible meaning.

From the NLP perspective, prediction of the unclear parts is performed through varying techniques and it usually utilizes linguistic information [4, 5]. For example, using N-grams is a popular uni-modal method for this task since they provide very robust predictions for local dependencies. Nevertheless, they lose their power for structures with long-range dependencies such as high-attached relative clauses, prepositional phrases, and repair utterances. Furthermore, if there are multiple instances of the same object class, which is a de facto case for daily environments (cf. Figure 1), N-grams cannot differentiate between them to select the proper instance reference. Alternatively, semantic clustering or classification (e.g. static word embeddings like word2vec [23]) are widely used for understanding the meaning communicated in speech. Recently, transformer-based deep learning approaches, that take the context (surrounding words) into account to efficiently capture the semantics of the word, have been successful in a set of NLP tasks, see [9] for the details of BERT embeddings, and [27, 33] for their use in multimodal settings. However, their success is also highly dependent on the size of the training data.

3 EYE-MOVEMENTS IN REFERENCE RESOLUTION

There is a considerable amount of literature on language, vision and their interaction (see [1] for a review). However, incorporating eye-movements in order to resolve ambiguities, especially in varying referential complexities, is still emerging area [12, 16–18, 24, 25].

Until recently, equipping robots with adequate eye-tracking devices was not a feasible solution due to their high cost. To mitigate this problem, one common approach was to use head-pose direction to roughly estimate the gaze direction. One study [34] uses a low-level visual saliency based method for establishing joint attention between an experimenter and a robot. This study addresses cases where there is no linguistic instruction or incorporation of high-level knowledge such as object labels. Instead, it aims to predict the targeted point via visual-saliency informed gaze direction from a low-resolution camera. Although their saliency method performs well for establishing joint attention, the authors also highlight the fact that head direction is not a substitute for gaze direction. However, as referenced in the previous sections, accurate eye-tracking becomes increasingly possible even with laptop cameras.

Another important aspect of incorporating eye-movements is the way of representing this modality. Henderson et al. [12] point out that the success of a system that identifies the attended objects is dependent on utilizing an effective combination of several fixation parameters instead of focusing on only one. This brings another parameter selection issue into foreground. Parameters like fixation location, duration or the gaze pattern are the main eye-movement metrics that have been widely used in gaze-contingent systems (e.g. [3, 31]). However, a lot of assumptions need to be made to decide when the aggregated group of eye-movements forms a fixation or saccade. These studies inspire us to represent eye-movements recorded during the comprehension of various scene–sentence pairs as a time-series feature vector to predict the attended object (see Section 5.3). Here, we encode a rich set of various raw fixation parameters without making any assumption about the type of eye-movements in a feature vector. Using this method, our ultimate goal is to get close to a gaze model that can generalize well for various sentence–scene (communication) complexities.

To conclude, with advancements in the eye-tracking technology, incorporating eye movements of a listener or speaker enables us to predict / resolve which entity is being referred to in a complex visual environment. However, it has also been shown that listener gaze can only be really beneficial when combined with situation-specific features of the current scene or language [17, 26]. Yet, those models are limited to reading activities or relatively simple scenes where the referential complexity is limited (due to language or visual clutter). Situated language understanding in a referentially complex or noisy environment imposes a different level of challenge due to uncertainty in the coupling among the modalities that, to the authors' knowledge, is still an uncharted area. Therefore, in addition to the contribution of eye-movements in language–vision

(a) HIGH-a: high-attached relative clause (b) HIGH-b: low-attached relative clause (c) MEDIUM: *fully visible target object*

Figure 2: A sample scene from the testing conditions (HIGH: *26 participants, 17 scenes*, MEDIUM: *27 participants, 19 scenes*)

Table 1: Sample sentences with varying complexities *(original experiment language is German)*, RC: Relative Clause

Complexity	Sample Sentence
1a.	It is a blanket on a sofa that she stitches.
High-attached RC	(*Da ist ein Tuch auf einer Couch die sie langsam stopft*)
1b.	It is a mug on a vitrine table that she damages.
Low-attached RC	(*Da befindet sich ein Becher auf einer Vitrine die sie achtlos beschädigt*)
2.	Hold the green blanket from the sofa.
No RC	(*Halte die grüne Decke vom Sofa*)

models, we also investigate how the coupling between between language and gaze is affected by referential complexity. This might give valuable insights to decide whether the referential complexity of the scene–sentence pair should be taken into account while deciding how to integrate a gaze-model.

4 APPLICATION SCENARIO

Augmented communication technologies are becoming part of our daily lives. Being able to follow a communication that conveys thoughts and intentions expressed in a flexible manner is a crucial component of NLP systems used for helper robots that aid people in their daily activities. Still, communication is not always noise-free. When the intention is not understood, the system can wait or ask for clarification. However, combining the uncertain information from the linguistic channel with information from the other ones increases the fluency and the effectiveness of the communication [10]. In this specific scenario, scene and gaze information is used to compensate the noise in the verbal channel.

Unlike standard NLP approaches, in this scenario, the helper robot is equipped with an eye-tracker so that the gaze of the speaker(s) can be tracked in an online fashion. Although the system is able to perform crossmodal mapping without incorporating the eye-movements by using SOTA NLP approaches, this modality is expected to improve the meaning recovery in noisy settings. Also, the use of the proposed method can be beneficial for other task-oriented communication scenarios, such as educational video games, training simulations and assistive navigation systems.

5 DATA

Eye4Ref [2] provides eye-movement recordings from various referentially complex situations defined at course granularity (collected with a *SR Eyelink 1000 Plus* eye tracker with a sampling rate of 1000 Hz). The data is coming from visual world paradigm experiments with a simple look-and-listen task, that presents participants with referentially complex images with accompanying spoken sentences. Referential complexity of the studies is controlled by visual and linguistic manipulations and quantified as HIGH, MEDIUM and LOW by simplistic approaches, such as the number of competing objects, or presence of ambiguities. Even though this is at course-grain level, in this study, we keep the original complexity classification to account for reproducibility. We employed two subsets of Eye4Ref dataset in this study; (i) HIGH referential complexity with ambiguous sentence in reference to a cluttered environment[1] and (ii) MEDIUM referential complexity with simple sentence also in reference to a cluttered environment[2], see Figure 2. The LOW condition was omitted from this study since it differs from the others in terms of the amount of visual clutter. In both conditions, for each mentioned object in the scene, there are distractor objects that share properties with the targets (e.g. same *type* or *color*). The HIGH condition also contains people and actions. To illustrate, in the first sentence in Table 1, while the target object is *blanket*, the set of other communicationally relevant objects (mentioned in the sentence) consists of *sofa* and *she*, which are all ambiguous (see Figure 2a).

Table 2: A partial list of the entities and scene representations in triplet notation for the scene in Figure 1.

	Entity Labels		Relations between Entities
1.	$robot_1$	1.	($woman_1$, AGENT, sit_1)
2.	$woman_1$	2.	($woman_2$, LOC_{NEXTTO}, $table_1$)
3.	$newspaper_1$	3.	($robot_1$, AGENT, $stand_1$)
4.	mug_1	4.	($robot_1$, LOC_{NEXTTO}, $sidetable_1$)
5.	$glasses_1$	5.	($newspaper_1$, LOC_{ON}, $tablecloth_1$)
6.	$sidetable_1$	6.	($newspaper_1$, LOC_{NEXTTO}, $woman_1$)
7.	$sidetable_2$	7.	($newspaper_1$, LOC_{NEXTTO}, mug_1)
8.	$sidetable_2$	8.	($glasses_1$, LOC_{ON}, $sidetable_1$)
9.	$glasses_2$	9.	($glasses_2$, LOC_{ON}, $sidetable_2$)
10.	$sidetable_1$	10.	($glasses_2$, COUNT, many)
11.	$kettle_1$	11.	(mug_1, COLOR, blue)

Eye4Ref also contains semantic scene representations. As illustrated in Table 2, all objects in the images are annotated as a triplet notation in the form `<entity, relation type, entity / property>`.

[1] 12 participants, 20 scene-sentence pairs and 240 eye-movement recordings
[2] 27 participants, 32 scene-sentence pairs in total 864 eye-movement recordings, please check [2] for the experimental details.

5.1 Masked sentences

In order to mimic noise in the language modality, we are masking words in the sentences. Regardless of whether they are the target object or not (*interest of action*), we treat all nouns and relative clause pronoun as masking candidates. Sentences in the HIGH condition have three candidates; the first noun (*blanket*), the second noun (*sofa*) and the RC pronoun (*she*) in Sentence 1a in Table 1. In the second study (MEDIUM, Sentence 2), there are only two candidates: *blanket* (first noun) and *sofa* (second noun). Each sentence has one masked word at a time, resulting in 51 sentence variations for HIGH and 38 variations for MEDIUM. Despite the small size of this test set, it should be noted that all multimodal pairs are unseen by both language and gaze models.

5.2 Gaze Features

We use a time-series format that requires less assumptions on the raw data. Time-series data also fit better for gaze-contingency since these systems need to make decisions incrementally without having any information after the present point in time. For computational efficiency, time-series eye-tracking data are usually analyzed in small bins. We use the provided scripts from the dataset repository to create bins with a cumulative sampling for 10 milliseconds.

The dataset contains recordings before sentence onset as a baseline, as well as the recordings while it is unfolding. We take the entire trial window for training and evaluation using the de-segmented fixation values for the overlapping objects provided by Eye4Ref dataset. To illustrate, lets assume that there is a *blanket* on the *side table*, and the area-of-interest (AOI) of the *side table* covers the AOI of the *blanket*. In such a case, if there is a fixation on the child object *blanket*, the raw data obviously also shows a fixation for the parent *side table*. Desegmentation removes fixations on the child objects from their parent objects.

5.3 Feature Vector

Eye4Ref provides pre-processed data for each scene and participant. For each sample (10 ms bin), all linguistic, scene and gaze features are represented in a feature vector. The size of the feature vector (on average 230 values) is dependent on the number of items in the scene. Approximately 180 dimensions correspond to one-hot encoded fixation location parameters[3] addressing all the objects in the scene. However, existing items are unique for each scene. Therefore, these fixated item features are not just sparse vectors but also at varying length. In our study, we reduce the size of this visual entity-specific feature vector to 2 features w.r.t. whether the gaze is (i) on the target object or (ii) on a communicationally relevant object. This gives us a fixed-length and task-specific feature vector that does not contain any location-related fixation information. The dimension of the final feature vector is 18, consisting of gaze and scene information; acceleration, velocity and their direction, angular resolution, pupil diameter, blink (or not), saccade (or not), binary parameters that signal whether the respective word (for *target* and all *comm. objects*) is being uttered at that point, and object count of the scene as a referential complexity measure. Acceleration, pupil size and velocity are normalized within participants and studies. In order to be able to generalize better, regardless of the object locations, gaze coordinates of the eye-movements are not included in the training since this information would be only useful in static images, where the objects have a fixed location.

[3]Binary values signifying whether the gaze is fixated on that object.

6 MODEL VARIATIONS

To establish the contribution of gaze features and understand their interaction with other communicational cues, we perform a study on multi-modal meaning recovery for noisy situations, where the linguistic modality is incomplete. For example, the word *blanket* is masked in the sentence "It is the … on the sofa that she silently stiches" and various LMs are tasked to fill the gap with a word (*task-1*) and to predict the corresponding entity in the scene (*task-2*). The full code is provided in the repository[4].

6.1 Bi-LSTM Model for the gaze processing

We employ a bi-directional LSTM architecture [13] using 50 LSTM nodes. For input, we create sequences of 50 samples, spanning 500 ms of input data. After the LSTM layer, we use two dense layers with 20 and 10 nodes respectively. For the binary classification on the single output layer, we use Sigmoid activation. Overall, the model contains 15,441 parameters[5].

We use a sequential machine learning setting to predict whether the gaze of the participant is on one of the communicationally relevant items while the spoken sentence unfolds. When utilizing language or rich scene representations, this task might be considered as trivial. However, when the multimodal setting involves complex verbal descriptions with ambiguous meanings or cluttered visual settings with object occlusion and multiple similar objects, the most salient modalities can fail. Then, eye-movements can contribute to resolving ambiguities and extracting meaningful predictions.

6.2 Language Models

As LMs, we employ several alternatives such as bi-directional N-grams from NLTK library [6] and pretrained contextualized embeddings like BERT, RoBERTa and XLM-RoBERTa masked language models [8, 9, 20] from the HuggingFace library [32].

The first N-gram model (Bi-N-gram(O)) is trained on the Eye4Ref dataset, where the test sentences also originate from. Order is set to 4 towards both directions, which makes it more sensitive to other tokens in the sentence. As we are training on the training data, this serves as a upper baseline. The other three models do not use Eye4Ref data during training. The second N-gram model was trained on HURIC 2.0 dataset [30], which contains commands for real human–robot interactions. This dataset contains 656 English sentences (vocabulary size: 481), addressing different situations representing possible commands given to a robot in a house environment. However – as the native language of the psycholinguistic experiments is German and the existing German resources fall short in covering human–robot interaction dialogues – we have translated HURIC 2.0 sentences into German using Google Translate Python library (vocabulary size: 449). We choose this dataset due to its resemblance to our setting, however it does not contain context

[4] https://gitlab.com/alacam/gaze2meaning

[5]Best parameters after grid search; Learning rate = 0.0001; Loss = binary crossentropy; Optimizer = Adam; Batch size = 200; Epochs = 100. The full code, models and predictions will be made available upon acceptance.

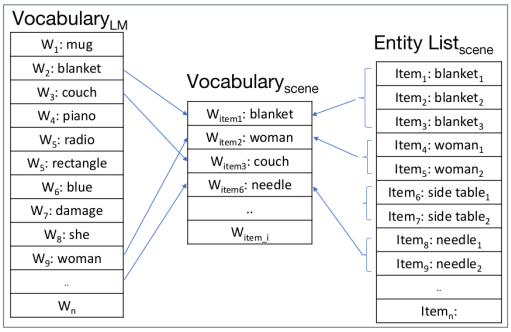


Figure 3: The vocabulary and entity lists

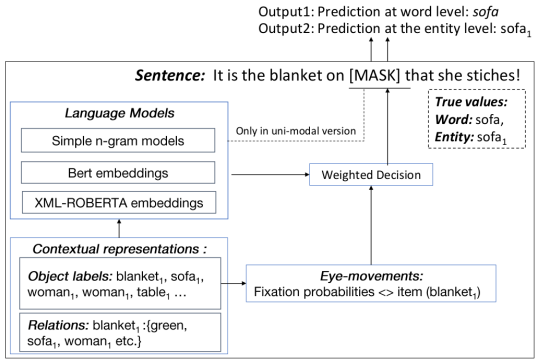


Figure 4: Multimodal ensemble for word/entity predictions

information of the scene or eye-movements, thus it is only utilized for training / fine-tuning the LMs.

As contextualized embeddings, we use various pre-trained BERT language models for German and English languages. All models were additionally fine-tuned on HURIC data. The performance of such models is highly dependent on the vocabulary size of the pretrained embeddings. Therefore, for each language, we used well-established large embeddings, which are pre-trained with SOTA methods. The German BERT Masked Language Model[6] has been trained on data from Wikipedia dumps, Open Subtitles, ParaCrawl and NewsCrawl resulting in a total dataset size of 16 GB and 2,350 million total tokens. For English, we use the RoBERTa Masked LM [20][7]. In addition, we also utilized `xlm-roberta-base` masked language model, which is a large multi-lingual language model, trained on 2.5 TB of filtered CommonCrawl data [8]. Unlike multi-lingual BERT and RoBERTa, XLM-RoBERTa can infer the language of the sentences. It is also chosen as an alternative to these more established models due to its SOTA scores for both languages.

To sum up, we have employed four LM variations for the word / entity prediction tasks for each language. For German, the models are bi-directional N-gram, bi-directional N-gram trained on Huric data, XLM-RoBERTa and dbmdz / GERMAN BERT models respectively. For English, we employ bi-directional N-gram, bi-directional N-gram trained on Huric data, XLM-RoBERTa and RoBERTa (large).

6.3 Tasks

6.3.1 Task-1: Word Recovery (finding the masked word). Assuming that we have complete semantic representations of the environment, the most straight-forward option would be to fill gaps based on utilizing rule-based inference. For example, the unknown reference has a LOCATION relation with a *sofa*, and either *sofa* or the *unknown item* has the THEME relation with the action *sew*. And there is only one object in the scene that has this action. By backtracking, we can predict that this entity must be the *blanket*$_1$, sewn by the *woman*$_1$, and located on the *sofa*$_1$. But the success of this method is highly dependent on the completeness of the extracted semantic representation, which is not usually the case. Therefore, we first utilize the power of language models such as skip-gram or CBOW models to fill the masked word based on the similarity and association scores [22]. The critical element of these models is the representatives of the training data / corpus, which might fall significantly short if the training samples are characteristically different than the testing ones. Our uni-modal LMs (L) are tasked to choose among their vocabulary (Vocabulary$_{LM}$) to predict the gap word as illustrated in Figure 3.

As indicated by [29], providing scene-related information is crucial for successful text completion in such settings. In the multi-modal versions, we provide language models with a list of candidate object labels and their properties in the scenes (scene representation). Based on this, the search space is reduced to objects and properties that exist in the scene, and among them, the most probable word or entity from the language model is selected as gap filler (*blanket*). Afterwards, we combine the LM with the gaze modality in an ensemble model, see Figure 4. To do this, the scores are converted into probabilities for each sentence and then best and top-k items are stored, as explained later.

6.3.2 Task-2: Entity Prediction. Once the word is predicted, the next step is to predict the entity being referred to. This process is done by filtering all the instances of the predicted word in the entity list, and then for each instances, we put its properties (e.g. color, shape) as well as the entities that the respective instance has a relation to into a list. We compare the sentence tokens against this list for each instance, and the entity with the highest similarity is selected as entity filler. While predicting the entity, we manipulated the way of incorporating scene information as *weak* and *strong* influence. In the weak visual influence condition, the entity prediction is highly dependent on the LM prediction since the arguments–sentence tokens comparison is performed only for the items that can be referred to by the predicted word. To illustrate, if the LM predicts a *blanket* as gap filler, the word *blanket* is searched in the scene items to find all candidates matching this prediction (e.g. blanket$_1$, blanket$_2$ or blanket$_3$). In the strong visual influence condition, the direction of the influence is the reversed; for each item in the scene, we calculate the LM score and the arguments–sentence tokens similarity score, and after normalization, we sum them in a linear way by using equal weights for each metrics.

In this study, we concentrate on achieving an exact match for the predicted item, Top-10 predictions are only used internally for the combination of the predictions of LMs and gaze models. However, all models can be adjusted to be evaluated for top-k recall and precision metrics (the prediction is classified as correct if the true label exists in the list of top-k predictions).

[6]https://huggingface.co/dbmdz/bert-base-german-cased
[7]https://huggingface.co/roberta-large

Table 3: Recall@1, 2, 3

	Word			Entity		
	R@1	**R@2**	**R@3**	**R@1**	**R@2**	**R@3**
Noun_1	53.68	84.21	94.74	51.16	78.95	84.21
Noun_2	42.11	57.89	63.16	26.32	52.63	63.16
Pronoun	31.58	47.37	68.42	15.79	36.84	52.63

7 RESULTS

7.1 Uni-modal and Context-informed LMs

Before combining various modalities in an ensemble model as described in Figure 4, we first aim to demonstrate the performance of uni-modal language models on such a situation-dependent meaning recovery task. The following sub-sections present the performance on German and translated English data.

German. Figure 5 presents the results of four LM variations for the word prediction task; bi-directional N-gram, bi-directional N-gram trained on HURIC data, XLM-RoBERTa and dbmdz/GERMAN BERT models respectively. As mentioned before, the N-gram (O) model serves as upper baseline, and provides valuable insights to understand the differences in masked position (Noun-1, Noun-2, or RC Pronoun). The poor results of the rest of the models indicate the difficulty of filling such gaps in the sentence for ambiguous cases by using only language (L) or context-informed language models (L+V). Despite their good performance on masked language tasks in general, where the context given in the surrounding text is incorporated, for the current task, situated context of the accompanying visual environment is more critical.

English. As shown in Figure 6, the models for the translated English sentences also exhibit similar patterns as in German experiments. The models are bi-directional N-gram, bi-directional N-gram trained on HURIC data, XLM-RoBERTa and ROBERTA (large). For the language models, the size of the vocabulary is one of the defining factors for higher performance, but for the current task, while finding the exact match for the word is one task, finding the exact entity in the referential word is the second and more challenging task, in which uni-modal or context-informed language models fail completely. The graphs consistently indicate that incorporating gaze information is crucial for the task at hand.

The prediction for the first masked position (Noun-1) is very low for uni-modal LMs, since the sentences have the same sentence structure and there are several other visual distractors that can be a masked candidate. In that case, it is very clear that gaze predictions boost the performance steadily. One interesting observation is that while German models are slightly better at predicting the Noun-1, the English models perform better at predicting the Noun-2. Further investigation is required to see whether this is related to the frequency of the words that fill these two positions or to the saliency of the objects. The masked words and the object labels usually share the same lexeme such as *blanket* and $blanket_1$, however due to the sentence structure, there is no direct match between the pronoun (s/he) and its entity in the scene ($woman_1$). The n-gram models do not have a multimodal co-reference resolution. Thus, it seems that even multimodal n-grams without scene or gaze support fail on filling the word and the entity for the pronoun position. BERT and XLM-RoBERTa language models do not seem to make this association better, indicating that a more elaborate solution is needed to address this issue.

7.2 Gaze Predictions

The gaze model provides a list of top 10 candidates with the probabilities for each possible masked position in the given sentence / scene setting. Table 3 summarizes the the recall scores of the target position being in the top-k of prediction list. Noun 1 is located in the highest rank for 53.7 % and 51.2 % of the cases for the word and entity respectively. And Recall@2 and Recall@3 values show a highly promising increase in performance, reaching 84.2 % and 94.7 % respectively. The recovery rate for the Noun-2 and RC. Pronoun is lower, especially for the entity prediction Recall@1. Considering that the gaze model is only trained on eye-movements without any guidance from the language or scene during training, this results draw a promising pattern.

7.3 Gaze-informed Language Models

In overall, the results indicate that gaze-informed language models performs significantly better than the unimodal LMs (except the upper baseline n-gram versions), Chi-square with Yates Correction $\chi^2(1) = 6.952$ (N=51, p < .05). Although the improvements for both Noun-1 and Noun-2 gap positions are clearly indicated, the L+V model and the L+V+G show similar performance overall (including the Pronoun position). This indicates that the recovery of concrete nouns like *blanket* or *sofa* is easier than recovering the personal pronoun (*s/he* or *it*) since the later one also requires indirect inference to the entity label $woman_i$ from the word *she*.

Interaction with Weak vs. Strong Context Influence. Figure 7 illustrates the interaction of gaze contribution with the context influence. The general trend indicates that the gaze-incorporated model with strong context influence performs better compared to the weak context influence. Therefore, for the sake of simplicity, we only report the models with the strong influence.

Gaze Model Variations. Eye-movements, as an indicator of where the attention is, are heavily influenced by the accompanying language and the characteristics of the visual environment. Therefore, the predictions from the gaze model were collected under three variations. The first one is based on simple fixation count (*no-learning*). It is calculated by summing up the fixations for each of object in the scene separately during the gap period as in:

$$F(item) = \sum_{i=onset}^{offset} F(item)_i$$

Figure 8 shows the prediction performance of the multimodal model with respect to three variations. In the second analysis, we use the predictions of the gaze model on the target object. Since spoken language unfolds in time and the focus of communication can be mentioned at any point (referenced or resolved later), having a one-to-one / tight alignment between the spoken word and the fixated object may not be always feasible as shown by the results. For a sanity check, we used the MEDIUM condition which does not have any semantic ambiguity. In such simple settings (like in Noun-2 in the MEDIUM condition, where there is no referential ambiguity), picking up the correct target object from the gaze model by only

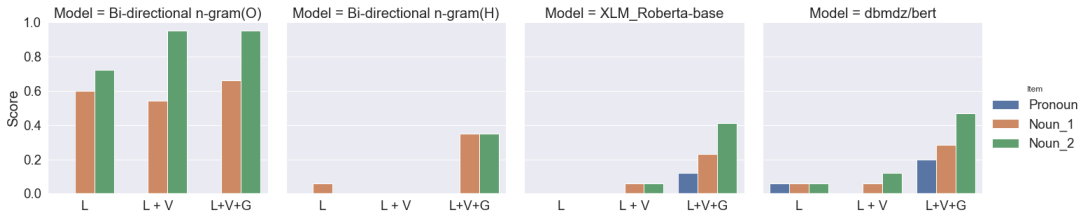
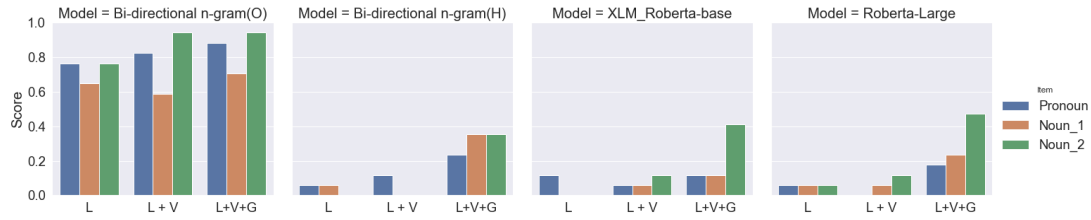
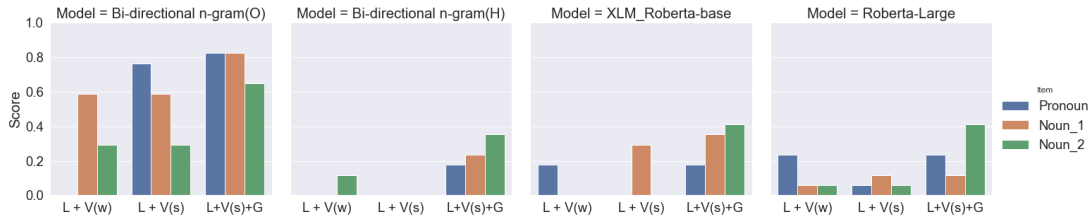


Figure 5: Predictions at the word level (German)



(a) Predictions at the word level (English)



(b) Predictions at the entity level (English)

Figure 6: F1-Scores (test: base model, test-tf: with transfer learning, test-tf: fine-tuned)

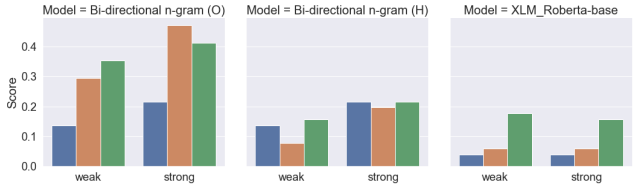


Figure 7: Gaze contribution: weak / strong context influence

focusing on the gap period is plausible. On the other hand, due to the referential ambiguity in the HIGH condition, where the ambiguity is resolved at the end of the sentence, less tighter alignment is required for successful prediction. To test this hypothesis, we compared the two periods; either only the gap period (*target period*) or the period of the *entire sentence*. For the former, we only check the correct predictions made while the target word is spoken, in the latter we extend it to an entire sentence period. In the HIGH condition, for the prediction of the entity referred to by the pronoun, no difference has been observed among these three variations *(no-learning, target and entire period)*. However, there is a clear advantage of using the entire period for the prediction of the second noun (*location object)*, while no consistent results are observed among different gaze-informed LMs performances for the topicalized noun (Noun-1). Still, pairwise comparison indicates that in overall using the entire period is better than the target period ($\chi^2(1) = 7.6645$, N=51, p < .05), while the target period and no learning conditions are very close to each other. For the MEDIUM condition, there are no significant differences between entire and target period predictions, and between no-learning and target period predictions (N=51, p > .05). Especially, for the second noun (no ambiguity), the entire period performs slightly worse than target and no-learning conditions indicates.

Weighted Predictions. Recall@2 and Recall@3 results indicate that the correct target entity is very competitive but due to common attention on the all mentioned objects they are not always ranked in the 1st position. For example, for overlapped objects like the mug_1 (Sentence-1b), the overlapping objects (despite gaze de-segmentation) often take precedence. Smaller objects can be viewed easily at the periphery vision once they are registered [21]. Therefore they are pushed further down in gaze predictions. But this might result in predictions like "it is a vitrine on the vitrine that she damages" which is highly unlikely. In order to punish such cases, we have reduced the weights of the all mentioned objects (vitrine_1, vitrine_2, woman_1, woman_2) by the factor of 0.5. They can still be selected in case the language model insists, but in general, this modification will allow possibly unmentioned and visually less salient entities to be selected. As seen in Figure 9, weighted gaze predictions result in better performance for all conditions. A pairwise comparison on the results of XLM-RoBERTa indicates that the weighted gaze predictions are significantly better than the original predictions ($\chi^2(1) = 6.195, N = 51, p < .05$), with Yates correction).

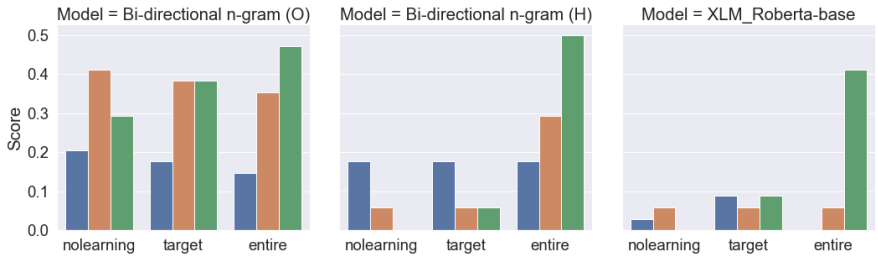


Figure 8: Gaze Model Variations: no learning, only target period, entire sentence

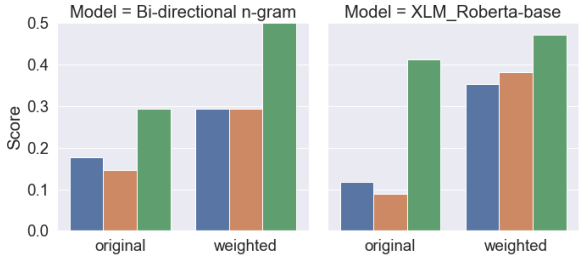


Figure 9: Weighted Predictions

8 CONCLUSION AND FUTURE WORK

In this study, we tested the contribution of the gaze modality integrated into SOTA NLP approaches for reference resolution from noisy settings. We have used the top-10 predictions of individual models and then combined them with a uniform linear weighting mechanism. The Recall@2 and Recall@3 scores also indicate that the relevant item, if not in the first position, still occurs in high ranks. Error analysis on misclassified cases indicates that the correct item for the masked period is often competing with other communicationally relevant objects. To exemplify, while the masked location is associated with woman$_1$, blanket$_1$ and sofa$_1$, which have been already uttered, receive higher attention. In order to mitigate this issue, enhancing the language models with part-of-speech tags, co-training the model on eye-movements alongside with the linguistic and visual information might allow the model to handle these intrinsic relations implicitly. But as a simple strategy, as shown here, punishing the entities of mentioned items by a weighting mechanism seems to perform well.

Despite the technological advancements in eye-tracking technologies, it might be still far-fetched to assume that the daily use of assistive technologies will be equipped with high-end eye-trackers in the near future. Our solution indicates that even highlighting the most communicationally relevant items, without establishing tight alignment between the spoken word and its reference in the scene, boosts the performance of a meaning recovery model. This kind of loose alignment can be easily achieved by low-frequency eye-trackers. Instead of focusing on an one-to-one coupling between gaze and the masked item, we demonstrated that distinguishing communicationally relevant objects is achievable based on the entire utterance period instead of extracting the most fixated item while the masked word is being spoken.

This analysis has been done in temporarily ambiguous and incomplete sentences accompanied by referentially complex environments (e.g. each mask candidate in the sentences has two other competitors). However, even in such linguistically and visually challenging cases, the models that incorporate gaze information consistently perform better on word and entity predictions compared to only Language (L) and Language&Visual (L+V) models. As can be seen in the Recall@1 scores in Table 3, the multimodal model (L+V+G) also performs better than the gaze-only model, which indicates that a) each modality can capture different aspects of the multimodal input, and b) their combination provides the best performance. In this study, we have shown that a combination of (off-the-shelf) masked LMs with a basic context integration and a gaze model can help to recover the meaning. This indicates that instead of developing complex language–vision ensemble models, a relatively simple gaze-incorporated multimodal integration might be feasible and computationally efficient. Furthermore, when the objects and their competitors are very close to each other, or when they are small, non-target objects seem to get substantial attention from the users. This, in very rare cases, impairs the overall performance of the multimodal ensemble although the LM has an accurate prediction for the target object. These issues require further investigation, which is highly dependent on the existence of large-scale multimodal datasets incorporating eye-movements.

We have used SOTA masked LMs in English and German, but they are also available for many languages. Besides, gaze-models can be considered as language-independent since we did not incorporate any linguistic features explicitly in the feature vector. The results indicate that transferring from German to English does not impair the results.

The next step is to evaluate the performance of the model in a simulated environment, where a virtual robot can react to speakers' commands and accompanying eye-movements. We also plan to record eye-movements in real-world settings to create gaze features and then systematically examine whether the gaze features can be transferred from one environment to another. These experiments will enable us to implement the integrated model on a service robot equipped with eye-tracking capabilities.

ACKNOWLEDGMENTS

This research was partially funded by the German Research Foundation DFG Transregio SFB 169: Cross-Modal Learning and by the Claussen-Simon-Foundation (funded project base.camp).

REFERENCES

[1] Özge Alaçam, Xingshan Li, Wolfgang Menzel, and Tobias Staron. 2020. Cross-modal Language Comprehension – Psycholinguistic Insights and Computational Approaches. *Frontiers in Neurorobotics* 14:2 (2020).

[2] Özge Alaçam, Eugen Ruppert, Amr R. Salama, Tobias Staron, and Wolfgang Menzel. 2020. Eye4Ref: A Multimodal Eye Movement Dataset of Referentially Complex Situations. In *Proceedings of the12th International Conference on Language Resources and Evaluation (LREC)*. Marseille, France, 2396–2404.

[3] Elena Arabadzhiyska, Okan Tarhan Tursun, Karol Myszkowski, Hans-Peter Seidel, and Piotr Didyk. 2017. Saccade landing position prediction for gaze-contingent rendering. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.

[4] Kavita Asnani, Douglas Vaz, Tanay PrabhuDesai, Surabhi Borgikar, Megha Bisht, Sharvari Bhosale, and Nikhil Balaji. 2015. Sentence Completion Using Text Prediction Systems. In *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014*. Springer, 397–404.

[5] Steffen Bickel, Peter Haider, and Tobias Scheffer. 2005. Learning to complete sentences. In *Proceedings of the 16th European Conference on Machine Learning*. Springer, Porto, Portugal, 497–504.

[6] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[7] Braiden Brousseau, Jonathan Rose, and Moshe Eizenman. 2020. Hybrid Eye-Tracking on a Smartphone with CNN Feature Extraction and an Infrared 3D Model. *Sensors* 20, 2 (2020), 543.

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv preprint arXiv:1301.3781* (2019).

[9] 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 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019)*. Minneapolis, Minnesota, USA, 4171–4186.

[10] Nestor Garay-Vitoria and Julio Abascal. 2004. A comparison of prediction techniques to enhance the communication rate. In *ERCIM Workshop on User Interfaces for All*. Springer, Vienna, Austria, 400–417.

[11] Edward Gibson, Leon Bergen, and Steven T. Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences* 110, 20, 8051–8056.

[12] John M. Henderson and Tim J. Smith. 2009. How are eye fixation durations controlled during scene viewing? Further evidence from a scene onset delay paradigm. *Visual Cognition* 17, 6-7 (2009), 1055–1082.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[14] Laszlo Hunyadi. 2013. Incompleteness and fragmentation in spoken language syntax and its relation to prosody and gesturing: Cognitive processes vs. Possible formal cues. In *IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*. Budapest, Hungary, 211–218.

[15] Mohamed Khamis, Florian Alt, and Andreas Bulling. 2018. The past, present, and future of gaze-enabled handheld mobile devices: survey and lessons learned. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*. Barcelona, Spain, 1–17.

[16] Sigrid Klerke and Barbara Plank. 2019. At a Glance: The Impact of Gaze Aggregation Views on Syntactic Tagging. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*. Hong Kong, China, 51–61.

[17] Nikolina Koleva, Martín Villalba, Maria Staudte, and Alexander Koller. 2015. The impact of listener gaze on predicting reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Vol. 2. Beijing, China, 812–817.

[18] Alexander Koller, Konstantina Garoufi, Maria Staudte, and Matthew Crocker. 2012. Enhancing referential success by tracking hearer gaze. In *Proceedings of the 13th annual meeting of the special interest group on discourse and dialogue*. Stroudsburg, PA, USA, 30–39.

[19] Roger Levy. 2008. A Noisy-channel Model of Rational Human Sentence Comprehension Under Uncertain Input. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*. Waikiki, Honolulu, Hawaii, 234–243.

[20] 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* (2019).

[21] Jack M. Loomis, Jonathan W. Kelly, Matthias Pusch, Jeremy N. Bailenson, and Andrew C Beall. 2008. Psychophysics of perceiving eye-gaze and head direction with peripheral vision: Implications for the dynamics of eye-gaze behavior. *Perception* 37, 9 (2008), 1443–1457.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, G. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. Lake Tahoe, Nevada, 3111–3119.

[24] Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada, 377–387.

[25] Nikolina Mitev, Patrick Renner, Thies Pfeiffer, and Maria Staudte. 2018. Using Listener Gaze to Refer in Installments Benefits Understanding. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*. Madison,Wisconsin, USA, 2122–2127.

[26] Zahar Prasov and Joyce Y Chai. 2008. What's in a gaze? The role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*. Gran Canaria, Spain, 20–29.

[27] Wasifur Rahman, Md Kamrul Hasan, Amir Zadeh, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 2019. M-BERT: Injecting Multimodal Information in the BERT Structure. *arXiv preprint arXiv:1908.05787* (2019).

[28] Sol Rogers. 2019. Seven Reasons Why Eye-tracking Will Fundamentally Change VR. *Forbes* (2019). https://www.forbes.com/sites/solrogers/2019/02/05/seven-reasons-why-eye-tracking-will-fundamentally-change-vr/

[29] Amr R. Salama, Özge Alaçam, and Wolfgang Menzel. 2018. Text Completion using Context-Integrated Dependency Parsing. In *Proceedings of the 3rd Workshop on Representation Learning for NLP - ACL 2018*. Melbourne, Australia, 41–49.

[30] Andrea Vanzo, Danilo Croce, Emanuele Bastianelli, Roberto Basili, and Daniele Nardi. 2020. Grounded language interpretation of robotic commands through structured learning. *Artificial Intelligence* 278 (2020).

[31] Cesco Willemse and Agnieszka Wykowska. 2019. In natural interaction with embodied robots, we prefer it when they follow our gaze: a gaze-contingent mobile eyetracking study. *Philosophical Transactions of the Royal Society B* 374, 1771 (2019), 20180036.

[32] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45.

[33] Jianfei Yu and Jing Jiang. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of the 28th International Joint Conferences on Artificial Intelligence*. Macao, China.

[34] Zeynep Yücel, Albert Ali Salah, Çetin Meriçli, Tekin Meriçli, Roberto Valenti, and Theo Gevers. 2013. Joint attention by gaze interpolation and saliency. *IEEE Transactions on cybernetics* 43, 3 (2013), 829–842.