Speaker-adapted neural-network-based fusion for multimodal reference resolution

Diana Kleingarn
Ruhr University Bochum
diana.kleingarn@rub.de

Martin Heckmann
Honda Research Institute Europe GmbH
martin.heckmann@honda-ri.de

Nima Nabizadeh
Ruhr University Bochum
nima.nabizadeh@rub.de

Dorothea Kolossa
Ruhr University Bochum
dorothea.kolossa@rub.de

Abstract

Humans use a variety of approaches to reference objects in the external world, including verbal descriptions, hand and head gestures, eye gaze or any combination of them. The amount of useful information from each modality, however, may vary depending on the specific person and on several other factors. For this reason, it is important to learn the correct combination of inputs for inferring the best-fitting reference. In this paper, we investigate speaker-dependent and independent fusion strategies in a multimodal reference resolution task. We show that without any change in the modality models, only through an optimized fusion technique, it is possible to reduce the error rate of the system on a reference resolution task by more than 50%.

1 Introduction

Reference resolution is of vital importance when human-machine interaction is expected to become natural and be integrated into everyday life. Humans have at their disposal a broad range of modalities to refer to objects in their environment, including verbal and material signals (Clark, 2005). Equipping machines with the capability to correctly interpret such reference resolutions raises the question of how to fuse the information derived from the different modalities.

Popular fusion methods in this domain can be categorized along two dimensions. The first is at which level of processing the fusion happens and the second how the fusion is performed (see Atrey et al. (2010); Ramachandram and Taylor (2017) for a comprehensive overview). In so-called early fusion or feature level fusion the features derived from the different modalities are combined, whereas in late fusion or decision level fusion classification results, e.g. in the form of probabilities, are combined. Regarding the second dimension, the methods are mainly grouped into classification-based and estimation-based methods.

As for the classification-based techniques, the modalities are usually combined at the feature level, i.e. early fusion, and the decision is obtained using a classifier. Iida et al. (2011) approached a reference resolution task in which two humans collaboratively solve a Tangram puzzle. Their method computed linguistic, gaze and task-specific features for each object of the board game and the objects were ranked using an SVM classifier. In a similar puzzle task, Funakoshi et al. (2012) proposed a model that could resolve verbal descriptions as well as gestures utilizing a Bayesian network. The Bayesian network design was later employed by Whitney et al. (2016) for interpreting referring expressions with speech and pointing gestures in a real-world cooking task.

Regarding the rule-based fusion, linear weighted fusion is one of the simplest and most widely used rule-based methods. This method combines the information from the different modalities linearly and it is assumed that the share of each modality in decision making does not change. It has been successfully utilized in multiple studies on reference resolution (Matuszek et al., 2014; Prasov and Yue Chai, 2010; Kennington et al., 2015; Kennington and Schlangen, 2017). A constraint-based rule system was used by (Holzapfel et al., 2004) where the constraints considered the time correlation of events and their semantic content for the fusion.

In this paper, we concentrate on one rule-based method used in Kennington et al. (2015). For this purpose, first, we explain the task and dataset in Sec. 2. Then, we discuss different approaches for the fusion of data in Sec. 3, including linear weighted fusion (Sec. 3.1) and our proposed neural-network-based fusion (Sec. 3.2), which

Proceedings of the SIGDial 2019 Conference, pages 210–214
Stockholm, Sweden, 11-13 September 2019. ©2019 Association for Computational Linguistics

also provides the possibility of learning speaker-dependent weights (Sec. 3.3). We summarize the results in Sec. 4 and give a short conclusion and outlook on future work in Sec. 5.

2 Previous work

2.1 The TAKE Dataset

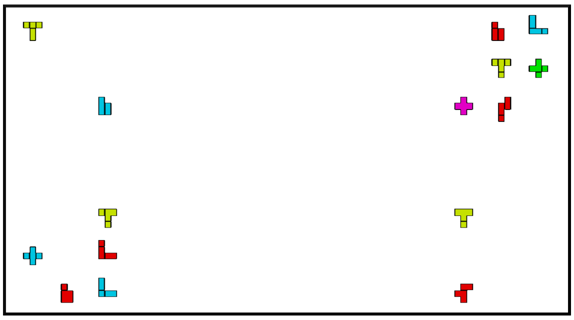

Figure 1: Example PENTO board on the TAKE dataset (Kennington and Schlangen, 2017)

The TAKE dataset was first introduced in Kousidis et al. (2013). It is a Wizard-of-Oz study, in which the participants were placed in front of a screen showing 15 pieces of a PENTO board game in random colors and shapes. The pieces were grouped into the four corners of the screen. For every episode, the shown objects and their positions on the screen was set randomly.

The participants were asked to instruct the system to select one specific PENTO piece on the board per episode. There was no instruction telling the participants how to refer to the item. According to the setup, it was possible to specify the object using spoken words, pointing gestures or eye gaze. Next, one piece was marked and the participant confirmed whether this selection was correct.

The example episode below, corresponding to Fig. 1, shows the English translation of the speech input and the true referent identifier:

- then we take now the se- so the second t that is on the top right ... out of this group there I would like to have the yellow t ... yes
- REFERENT o3

For this work, the confirmation utterance, e.g. the word “yes” in the above example, was removed, since it is not available at the time the decision is made. After this cleanup, the dataset includes 1034 episodes distributed over 7 users as shown in Table 1. The participants were native speakers, except for one, who spoke proficient but not native German.

User	Episodes	With pointing	With gaze
1	90	87	71
2	66	29	64
3	133	35	126
4	230	209	212
5	146	13	130
6	176	78	157
7	193	162	164
Total	1034	613	924

Table 1: Number of episodes, per user and cumulatively, in the TAKE dataset.

The speech, an average of 6.8 words per utterance, was transcribed using Google Web Speech as an automatic speech recognition (ASR), with a vocabulary size of 1049. Additionally, the speech was transcribed by hand, which can provide a reasonable upper bound for the results. A Microsoft Kinect above the screen captured the arm movements and an eye tracker (*Seeingmachines FaceLab*) was used to determine the eye gaze.

Since the scenes in this dataset are virtual, we can directly annotate the objects with the properties and then query the scene representation. For this simplified task, the properties are the color, the shape and the spatial relations of the pieces. Using image processing techniques described in Kennington et al. (2015), several features for each object are extracted, including the number of edges, RGB (red, green, blue) values, HSV (hue, saturation, value), its centroid, horizontal and vertical skewness, and the orientation value denoting the direction of the principal axis. These features are used for the natural language grounding described in the next section.

2.2 Model for Natural Language Understanding

The idea is to treat each word in the vocabulary as a classifier which can relate the word to the perceptual information of the objects. For this purpose, a logistic regression classifier is trained to map the visual features $\mathbf{x}$ of each particular candidate object to a probability p_w of these features, given the word w.

$$p_w(\mathbf{x}) = \sigma(\mathbf{w}^\intercal \mathbf{x} + b) \tag{1}$$

Here, $\mathbf{w}$ is the learned weight vector and σ is the logistic function. What is needed for further steps, however, is one distribution over all candidate objects per episode. To accomplish that, we can average the distribution of all time steps $n = 1 \ldots N$

and normalize the prediction score of each object ($i \in I$) over all the $|I| = 15$ object candidates via

$$p_{\text{speech}}(i) = \frac{\sum_{n=1}^{N} p_{w_n}(\mathbf{x}_i)}{\sum_{k=1}^{|I|} \sum_{n=1}^{N} p_{w_n}(\mathbf{x}_k)} . \tag{2}$$

2.3 Model for Pointing Gestures and Gaze

For gaze and pointing gestures, we need a model that takes the coordinates of gaze and pointing as its input and returns a probability distribution over the object candidates, given the location of objects. This model is the same for gaze and pointing gestures.

For this purpose, we compute the average of the gaze or pointing coordinates for each episode, producing a reference point (R) for the modality. The reference point is compared to the centroid of each object (x_i, y_i) using a Gaussian distribution,

$$p_{\text{d}}(i) \propto \exp\left[-\frac{(x_R - x_i)^2}{2 \cdot \sigma_x^2} - \frac{(y_R - y_i)^2}{2 \cdot \sigma_y^2}\right] . \tag{3}$$

The result is then normalized over all objects to obtain p_{point} and p_{gaze}, so that the objects closer to the reference point will have a higher probability.

3 Fusion Models

3.1 Linear Fusion

For optimum performance, all three modalities need to be combined. A simple approach is to perform a rule-based late fusion by estimating a fixed weight for each modality and then summing the weighted prediction distributions, as in Kennington (2016):

$$\begin{aligned} p(i) = p_{\text{speech}}(i) \cdot \alpha_1 + p_{\text{point}}(i) \cdot \alpha_2 \\ + p_{\text{gaze}}(i) \cdot (1 - \alpha_1 - \alpha_2) . \end{aligned} \tag{4}$$

The system then makes a maximum-likelihood decision according to

$$\hat{i} = \arg\max_{i \in I} p(i) . \tag{5}$$

3.2 Neural-network-based Fusion

In Sec. 3.1, a baseline approach to late fusion is shown. To decrease the error rate, we now propose a more flexible method, which can model non-linear relations between the modalities. For this purpose we chose a fully connected neural network with one hidden layer, 512 neurons and a rectified linear unit as the activation function.

Its inputs $\mathbf{o}$ are the three concatenated modality vectors from (2) and (3),

$$\begin{aligned} \mathbf{o} &= [\mathbf{p}_{\text{speech}}, \mathbf{p}_{\text{point}}, \mathbf{p}_{\text{gaze}}] \\ \text{with} \quad \mathbf{p} &= [p_1, \ldots, p_{|I|}]. \end{aligned} \tag{6}$$

The output layer uses the softmax function so that the output can be interpreted as a probability distribution and used in Eq. (5) to obtain the estimated referent. To optimize the network parameters, we carried out preliminary tests with differently sized hidden layers and with additional reliability information, e.g., the variance of gaze or pointing information. For hand-annotated data, including the variance of all deixis coordinates of the current episode, $\mathbf{V}$, in the observation vector gave the best results. With this update, the network input becomes

$$\mathbf{o} = [\mathbf{p}_{\text{speech}}, \mathbf{p}_{\text{point}}, \mathbf{p}_{\text{gaze}}, \mathbf{V}]. \tag{7}$$

3.3 Speaker adaptation

Humans have different preferences in the way they refer to objects. This is also reflected in the dataset, in which many episodes from one participant are quite alike, whereas significant differences can often be observed across participants. Hence, depending on the participant, different modalities are very likely to contribute a variable amount of useful information. A model that adapts to a specific user should therefore outperform a general model.

However, judging from the small number of samples per user in Tab. 1, it is evidently not promising to train a neural network using only the data of one participant. Inspired by Saon et al. (2013), we addressed this problem by training on the full training set and reducing to a smaller training set, containing just one user, for the last 5 % of the epochs.

4 Evaluation

We evaluate all fusion methods on the same data as Kennington et al. (2015) under the same four conditions: speech only, speech with gaze, speech with deixis, and speech with gaze and deixis. For this purpose, we compare the error rate $E = 100 \cdot \frac{M-C}{M}$ under all conditions, with C as the number of correctly estimated referents, among M estimates made for the test set.

However, for the linear fusion with fixed weights (fw) presented in Sec. 3.1, we did not use the weights suggested in Kennington et al. (2015). Instead, a grid search was run on the training data to determine optimal weights for the dataset (ow). This yielded an average improvement of 5.9% absolute for hand-annotated data and also improved all individual cases for ASR-annotated data except for the fusion of all modalities. Here, the results slightly deteriorated from 60.3% to 60.0%.

We used 10-fold cross validation to obtain an estimate of the error rate together with its standard deviation. These results are depicted in Fig. 2. As can be seen, there is a large difference in performance between the results using the hand-annotated speech data vs. the ASR system, indicating a likely high number of transcription errors for the informative keywords. It can also be seen that adding more modalities consistently improves the performance.

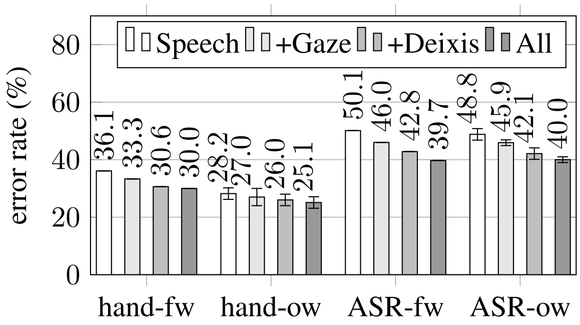

Figure 2: Error rate (%) and standard deviation for optimized (ow) or fixed weights (fw, adapted from (Kennington et al., 2015)) in (4).

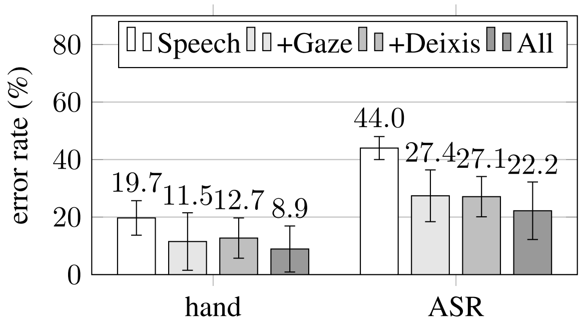

Figure 3: Error rate (%) of the proposed neural-network-based fusion

The neural network-based fusion (Sec. 3.2) increased performance compared to the linear fusion (fw) notably and for all conditions. These results are shown in Fig. 3. We obtain the best results with an error rate of 8.9% for the fusion of all modalities using the hand-annotated data. In comparison to the fixed-weight baseline, with an error rate of 30% (see Fig. 2), the error rate is hence decreased by 70%.

User	NN ASR	NN hand
User 1	17.5 (±8.5)	3.4 (±5.8)
	35.8 (±10.9)	8.5 (±5.6)
User 2	11.7 (±13.9)	11.0 (±11.9)
	16.2 (±8.9)	12.1 (±11.6)
User 3	10.8 (±12.8)	3.5 (±6.5)
	11.9 (±11.8)	4.5 (±8.7)
User 4	12.3 (±10.2)	5.7 (±6.2)
	10.5 (±9.1)	5.2 (±6.9)
User 5	22.6 (±7.9)	6.5 (±7.7)
	28.3 (±11.8)	11.3 (±8.9)
User 6	19.1 (±8.8)	12.5 (±8.9)
	23.4 (±10.6)	14.8 (±12.5)
User 7	31.0 (±13.5)	6.4 (±9.0)
	24.0 (±10.7)	4.7 (±7.0)
average	18.6 (±10.7)	7.0 (±7.8)
	22.2 (±10.3)	8.9 (±8.2)

Table 2: Results of the user-dependent (black) and the user-independent (gray) model in terms of error rate (%) and standard deviation σ.

Table 2 compares the results of the speaker-dependent and -independent models for each user. Here, we only report the results for the fusion of all modalities. When using the hand annotation, the speaker-adapted fusion reduces the error rate further, from 8.9% to 7.0%. But it can also be seen that the results vary largely from user to user. In particular, for user 1 (ASR data), the speaker-adapted version outperforms the other version easily, but for user 7, the original, speaker-independent version is more accurate. For hand-annotated data, the difference between the two versions is smaller, but the users for which the speaker-adapted version outperforms the other remain the same. Interestingly the speaker-adapted version performs least well for the two users with the most episodes that mostly contain gaze and pointing information, as can be seen in Table 1.

5 Conclusions

We have compared different fusion strategies for multi-modal information integration in a reference resolution task. Our results show that a fully connected neural network can reduce the error rate significantly, compared to a weighted averaging of single-modality posterior probabilities. Adapting the fusion to each specific user is also helpful to some extent, although the improvements are less clear and consistent.

In this work, we applied fairly simple models for speech, gaze and pointing, which simply use the average values of all features for the current episode. Since some words carry more semantic content than others for finding the referent, and since the coordinate sequences of gaze and pointing contain some redundancy, as well as segments of more and of less information content, future work will focus on the creation of a time-dependent model for improving multi-modal fusion.

6 Acknowledgments

We want to thank Casey Kennington for fruitful discussions and his support in setting up the model for linear fusion. Furthermore, we are grateful to Casey Kennington and Sina Zarrie for providing the data. In addition, many thanks to Julia Ringeis for supporting us in preparing the manuscript.

References

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379.

Herbert H Clark. 2005. Coordinating with each other in a material world. *Discourse studies*, 7(4-5):507–525.

Kotaro Funakoshi, Mikio Nakano, Takenobu Tokunaga, and Ryu Iida. 2012. A unified probabilistic approach to referring expressions. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2012)*, pages 237–246.

Hartwig Holzapfel, Kai Nickel, and Rainer Stiefelhagen. 2004. Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gestures. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 175–182.

Ryu Iida, Masaaki Yasuhara, and Takenobu Tokunaga. 2011. Multi-modal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 84–92.

Casey Kennington, Livia Dia, and David Schlangen. 2015. A Discriminative Model for Perceptually-Grounded Incremental Reference Resolution. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS) 2015*, pages 195–205.

Casey Kennington and David Schlangen. 2017. A Simple Generative Model of Incremental Reference Resolution for Situated Dialogue. *Comput. Speech Lang.*, 41(C):43–67.

Casey Redd Kennington. 2016. *Incrementally Resolving References in Order to Identify Visually Present Objects in a Situated Dialogue Setting*. Ph.D. thesis, Bielefeld University.

Spyros Kousidis, Casey Kennington, and David Schlangen. 2013. Investigating speaker gaze and pointing behaviour in human-computer interaction with the mint. tools collection. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2013)*, pages 319–323.

Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 2556–2563.

Zahar Prasov and Joyce Yue Chai. 2010. Fusing eye gaze with speech recognition hypotheses to resolve exophoric references in situated dialogue. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 471–481.

Dhanesh Ramachandram and Graham W Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108.

George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny. 2013. Speaker adaptation of neural network acoustic models using i-vectors. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 55–59.

David Whitney, Miles Eldon, John Oberlin, and Stefanie Tellex. 2016. Interpreting multimodal referring expressions in real time. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3331–3338.