Multimodal Semantic Learning from Child-Directed Input

Angeliki Lazaridou
University of Trento
angeliki.lazaridou@unitn.it

Grzegorz Chrupała
Tilburg University
g.chrupala@uvt.nl

Raquel Fernández
University of Amsterdam
raquel.fernandez@uva.nl

Marco Baroni
University of Trento
marco.baroni@unitn.it

Abstract

Children learn the meaning of words by being exposed to perceptually rich situations (linguistic discourse, visual scenes, etc). Current computational learning models typically simulate these rich situations through impoverished symbolic approximations. In this work, we present a *distributed* word learning model that operates on child-directed speech paired with realistic *visual scenes*. The model integrates linguistic and extra-linguistic information (visual and social cues), handles referential uncertainty, and correctly learns to associate words with objects, even in cases of limited linguistic exposure.

1 Introduction

Computational models of word learning typically approximate the perceptual context that learners are exposed to through artificial proxies, e.g., representing a visual scene via a collection of symbols such as `cat` and `dog`, signaling the presence of a cat, a dog, etc. (Yu and Ballard, 2007; Fazly et al., 2010, *inter alia*).[1] While large amounts of data can be generated in this way, they will not display the complexity and richness of the signal found in the natural environment a child is exposed to. We take a step towards a more realistic setup by introducing a model that operates on naturalistic images of the objects present in a communicative episode. Inspired by recent computational models of meaning (Bruni et al., 2014; Kiros et al., 2014; Silberer and Lapata, 2014), that integrate distributed linguistic and visual information, we build upon the Multimodal Skip-Gram (MSG) model of Lazaridou et al. (2015). and enhance it to handle cross-referential uncertainty. Moreover, we extend the cues commonly used in multimodal learning (e.g., objects in the environment) to include *social cues* (e.g., eye-gaze, gestures, body posture, etc.) that reflect speakers' intentions and generally contribute to the unfolding of the communicative situation (Stivers and Sidnell, 2005). As a first step towards developing full-fleged learning systems that leverage all signals available within a communicative setup, in our extended model we incorporate information regarding the objects that caregivers are holding.

[1] See Kádár et al. (2015) for a recent review of this line of work, and another learning model using, like ours, real visual input.

2 Attentive Social MSG Model

Like the original MSG, our model learns multimodal word embeddings by reading an utterance sequentially and making, for each word, two sets of predictions: (a) the preceding and following words, and (b) the visual representations of objects co-occurring with the utterance. However, unlike Lazaridou et al. (2015), we do not assume we know the right object to be associated with a word. We consider instead a more realistic scenario where multiple words in an utterance co-occur with multiple objects in the corresponding scene. Under this *referential uncertainty*, the model needs to induce word-object associations as part of learning, relying on current knowledge about word-object affinities as well as on any social clues present in the scene.

Similar to the standard skipgram, the model's parameters are context word embeddings $\mathbf{W}'$ and tar-

get word embeddings $\mathbf{W}$. The model aims at optimizing these parameters with respect to the following multi-task loss function for an utterance w with associated set of objects U:

$$L(w, U) = \sum_{t=1}^{T} (\ell_{ling}(w, t) + \ell_{vis}(w_t, U)) \quad (1)$$

where t ranges over the positions in the utterance w, such that w_t is t^{th} word. The linguistic loss function is the standard skip-gram loss (Mikolov et al., 2013). The visual loss is defined as:

$$\ell_{vis}(w_t, U) = \sum_{s=1}^{S} \lambda \alpha(\mathbf{w}_t, \mathbf{u}_s) g(\mathbf{w}_t, \mathbf{u}_s) + (1 - \lambda) h(\mathbf{u}_s) g(\mathbf{w}_t, \mathbf{u}_s) \quad (2)$$

where $\mathbf{w}_t$ stands for the column of $\mathbf{W}$ corresponding to word w_t, $\mathbf{u}_s$ is the vector associated with object U_s, and g the penalty function

$$g(\mathbf{w}_t, \mathbf{u}_s) = \sum_{\mathbf{u}'} \max(0, \gamma - \cos(\mathbf{w}_t, \mathbf{u}_s) + \cos(\mathbf{w}_t, \mathbf{u}')), \quad (3)$$

which is small when projections to the visual space $\mathbf{w}_t$ of words from the utterance are similar to the vectors representing co-occurring objects, and at the same time they are dissimilar to vectors $\mathbf{u}'$ representing randomly sampled objects. The first term in Eq. 2 is the penalty g weighted by the current word-object affinity α, inspired by the "attention" of Bahdanau et al. (2015). If α is set to a constant 1, the model treats all words in an utterance as equally relevant for each object. Alternatively it can be used to encourage the model to place more weight on words which it already knows are likely to be related to a given object, by defining it as the (exponentiated) cosine similarity between word and object normalized over all words in the utterance:

$$\alpha(\mathbf{w}_t, \mathbf{u}_s) = \frac{\exp(\cos(\mathbf{w}_t, \mathbf{u}_s))}{\sum_r \exp(\cos(\mathbf{w}_r, \mathbf{u}_s))} \quad (4)$$

The second term of Eq. 2 is the penalty weighted by the social salience h of the object, which could be based on various cues in the scene. In our experiments we set it to 1 if the caregiver holds the object, 0 otherwise.

We experiment with three versions of the model. With $\lambda = 1$ and α frozen to 1, the model reduces to the original **MSG**, but now trained with referential uncertainty. The **Attentive MSG** sets $\lambda = 1$ and calculates $\alpha(\mathbf{w}_t, \mathbf{u}_s)$ using Equation 4 (we use the term "attentive" to emphasize the fact that, when processing a word, the model will pay more attention to the more relevant objects). Finally, **Attentive Social MSG** further sets $\lambda = \frac{1}{2}$, boosting the importance of socially salient objects.

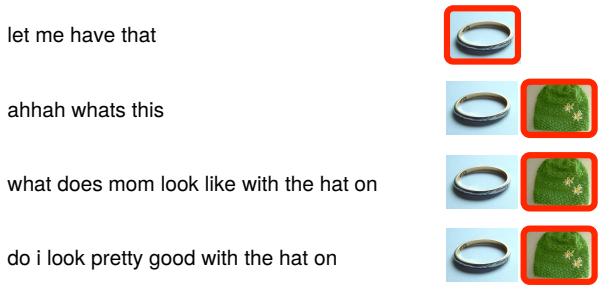

Figure 1: Fragment of the IFC corpus where symbolic labels `ring` and `hat` have been replaced by real images. Red frames mark objects being touched by the caregiver.

All other hyperparameters are set to the values found by Lazaridou et al. (2015) to be optimal after tuning, except hidden layer size that we set to 200 instead of 300 due to the small corpus (see Section 3). We train the MSG models with stochastic gradient descent for one epoch.

3 The Illustrated Frank et al. Corpus

Frank et al. (2007) present a Bayesian cross-situational learning model for simulating early word learning in first language acquisition. The model is tested on a portion of the Rollins section of the CHILDES Database (MacWhinney, 2000) consisting of two transcribed video files (`me03` and `di06`), of approximately 10 minutes each, where a mother and a pre-verbal infant play with a set of toys. By inspecting the video recordings, the authors manually annotated each utterance in the transcripts with a list of object labels (e.g., `ring`, `hat`, `cow`) corresponding to all midsize objects judged to be visible to the infant while the utterance took place, as well as various social cues. The dataset includes a gold-standard lexicon consisting of 36 words paired with 17 object labels (e.g., *hat*=`hat`, *pig*=`pig`, *piggie*=`pig`).[2]

[2] `http://langcog.stanford.edu/materials/nipsmaterials.html`

Aiming at creating a more realistic version of the original dataset, akin to simulating a real visual scene, we replaced symbolic object labels with actual visual representations of objects. To construct such visual representations, we sample for each object 100 images from the respective ImageNet (Deng et al., 2009) entry, and from each image we extract a 4096-dimensional visual vector using the Caffe toolkit (Jia et al., 2014), together with the pre-trained convolutional neural network of Krizhevsky et al. (2012).[3] These vectors are finally averaged to obtain a single visual representation of each object. Concerning social cues, since infants rarely follow the caregivers' eye gaze but rather attend to objects held by them (Yu and Smith, 2013), we include in our corpus only information on whether the caregiver is holding any of the objects present in the scene. Note however that this signal, while informative, can also be ambiguous or even misleading with respect to the actual referents of a statement. Figure 1 exemplifies our version of the corpus, the Illustrated Frank et al. Corpus (IFC).

Several aspects make IFC a challenging dataset. Firstly, we are dealing with language produced in an interactive setting rather than written discourse. For example, compare the first sentence in the Wikipedia entry for *hat* (*"A hat is a head covering"*) to the third utterance in Figure 1, corresponding to the first occurrence of *hat* in our corpus. Secondly, there is a large amount of referential uncertainty, with up to 7 objects present per utterance (2 on average) and with only 33% of utterances explicitly including a word directly associated with a possible referent (i.e., not taking into account pronouns). For instance, the first, second and last utterances in Figure 1 do not explicitly mention any of the objects present in the scene. This uncertainty also extends to social cues: only in 23% of utterances does the mother explicitly name an object that she is holding in her hands. Finally, models must induce word–object associations from minimal exposure to input rather than from large amounts of training data. Indeed, the IFC is extremely small by any standards: 624 utterances making up 2,533 words in total, with 8/37 test words occurring only once.

[3]To match the hidden layer size, we average every k = 4096/200 original non-overlapping visual dimensions into a single dimension.

Model	*Best-F*
MSG	.64 (.04)
AttentiveMSG	.70 (.04)
AttentiveSocialMSG	.73 (.03)
ASMSG+shuffled visual vectors	.65 (.06)
ASMSG+randomized sentences	.59 (.03)
BEAGLE	.55
PMI	.53
Bayesian CSL	.54
BEAGLE+PMI	.83

Table 1: Best-F results for the MSG variations and alternative models on word-object matching. For all MSG models, we report Best-F mean and standard deviation over 100 iterations.

4 Experiments

We follow the evaluation protocol of Frank et al. (2007) and Kievit-Kylar et al. (2013). Given 37 test words and the corresponding 17 objects (see Table 2), all found in the corpus, we rank the objects with respect to each word. A mean *Best-F* score is then derived by computing, for each word, the top F score across the precision-recall curve, and averaging it across the words. MSG rankings are obtained by directly ordering the visual representations of the objects by cosine similarity to the MSG word vectors.

Table 1 reports our results compared to those in earlier studies, all of which did not use actual visual representations of objects but rather arbitrary symbolic IDs. Bayesian CSL is the original Bayesian cross-situational model of Frank et al. (2007), also including social cues (not limited, like us, to mother's touch). BEAGLE is the best semantic-space result across a range of distributional models and word-object matching methods from Kievit-Kylar et al. (2013). Their distributional models were trained in a *batch mode*, and by treating object IDs as words so that standard word-vector-based similarity methods could be used to rank objects with respect to words. Plain MSG is outperforming nearly all earlier approaches by a large margin. The only method bettering it is the BEAGLE+PMI combination of Kievit-Kylar et al. (PMI measures direct co-occurrence of test words and object IDs). The latter was obtained through a grid search of all possible model combinations performed directly on the test set, and relied on a weight parameter optimized on the corpus by assuming access to gold annotation.

It is thus not comparable to the untuned MSG.

Plain MSG, then, performs remarkably well, even without any mechanism attempting to track word-object matching across scenes. Still, letting the model pay more attention to the objects currently most tightly associated to a word (AttentiveMSG) brings a large improvement over plain MSG, and a further improvement is brought about by giving more weight to objects touched by the mother (AttentiveSocialMSG). As concrete examples, plain MSG associated the word *cow* with a pig, whereas AttentiveMSG correctly shifts attention to the cow. In turn, AttentiveSocialMSG associates to the right object several words that AttentiveMSG wrongly pairs with the hand holding them, instead.

One might fear the better performance of our models might be due to the skip-gram method being superior to the older distributional semantic approaches tested by Kievit-Kylar et al. (2013), independently of the extra visual information we exploit. In other words, it could be that MSG has simply learned to treat, say, the *lamb* visual vector as an arbitrary signature, functioning as a semantically opaque ID for the relevant object, without exploiting the visual resemblance between lamb and sheep. In this case, we should obtain similar performance when arbitrarily shuffling the visual vectors across object types (e.g., consistently replacing each occurrence of the *lamb* visual vector with, say, the *hand* visual vector). The lower results obtained in this control condition (ASMSG+shuffled visual vector) confirm that our performance boost is largely due to exploitation of genuine visual information.

Since our approach is incremental (unlike the vast majority of traditional distributional models that operate on batch mode), it can in principle exploit the fact that the linguistic and visual flows in the corpus are meaningfully ordered (discourse and visual environment will evolve in a coherent manner: a hat appears on the scene, it's there for a while, in the meantime a few statements about hats are uttered, etc.). The dramatic quality drop in the ASMSG+randomized sentences condition, where AttentiveSocialMSG was trained on IFC after randomizing sentence order, confirms the coherent situation flow is crucial to our good performance.

word	*gold object*	*17 objects*		*5.1K objects*	
		nearest	*r*	*nearest*	*r*
bunny	bunny	bunny	**1**	bunny	**1**
cows	cow	cow	**1**	lea	**7**
duck	duck	duck	**1**	mallard	**4**
duckie	duck	duck	**1**	mallard	**3**
kitty	kitty	book	**2**	bookcase	**66**
lambie	lamb	lamb	**1**	lamb	**1**
moocows	cow	cow	**1**	ranch	**4**
rattle	rattle	rattle	**1**	rattle	**1**

Table 2: Test words occurring only once in IFC, together with corresponding gold objects, AttentiveSocialMSG top visual neighbours among the test items and in a larger 5.1K-objects set, and ranks of gold object in the two confusion sets.

Minimal exposure. Given the small size of the input corpus, good performance on the word-object association already counts as indirect evidence that MSG, like children, can learn from small amounts of data. In Table 2 we take a more specific look at this challenge by reporting AttentiveSocialMSG performance on the task of ranking object visual representations for test words that occurred *only once* in IFC, considering both the standard evaluation set and a much larger confusion set including visual vectors for 5.1K distinct objects (those of Lazaridou et al. (2015)). Remarkably, in all but one case, the model associates the test word to the right object from the small set, and to either the right object or another relevant visual concept (e.g., a ranch for *moocows*) when the extended set is considered. The exception is *kitty*, and even for this word the model ranks the correct object as second in the smaller set, and well above chance for the larger one. Our approach, just like humans (Trueswell et al., 2013), can often get a word meaning right based on a single exposure to it.

Generalization. Unlike the earlier models relying on arbitrary IDs, our model is learning to associate words to actual feature-based visual representations. Thus, once the model is trained on IFC, we can test its generalization capabilities to associate known words with new object instances that belong to the right category. We focus on 19 words in our test set corresponding to objects that were normed for visual similarity to other objects by Silberer and Lapata (2014). Each test word was paired with 40 ImageNet pictures evenly divided between images of the gold

object (*not* used in IFC), of a highly visually similar object, of a mildly visually similar object and of a dissimilar one (for *duck*: *duck*, *chicken*, *finch* and *garage*, respectively). The pictures were represented by vectors obtained with the same method outlined in Section 3, and were ranked by similarity to a test word AttentiveSocialMSG representation.

Average Precision@10 for retrieving gold object instances is at 62% (chance: 25%). In the majority of cases the top-10 intruders are instances of the most visually related concepts (60% of intruders, vs. 33% expected by chance). For example, the model retrieves pictures of sheep for the word *lamb*, or bulls for *cow*. Intriguingly, this points to classic overextension errors that are commonly reported in child language acquisition (Rescorla, 1980).

5 Related Work

While there is work on learning from multimodal data (Roy, 2000; Yu, 2005, a.o.) as well as work on learning distributed representations from child-directed speech (Baroni et al., 2007; Kievit-Kylar and Jones, 2011, a.o.), to the best of our knowledge ours is the first method which learns distributed representations from multimodal child-directed data. For example, in comparison to Yu (2005)'s model, our approach (1) induces distributed representations for words, based on linguistic and visual context, and (2) operates entirely on distributed representations through similarity measures without positing a categorical level on which to learn word-symbol/category-symbol associations. This leads to rich multimodal conceptual representations of words in terms of distributed multimodal features, while in Yu's approach words are simply distributions over categories. It is therefore not clear how Yu's approach could capture phenomena such as predicting appearance from a verbal description or representing abstract words–all tasks that our model is at least in principle well-suited for. Note also that Frank et al. (2007)'s Bayesian model we compare against could be extended to include realistic visual data in a similar vein to Yu's, but it would then have the same limitations.

Our work is also related to research on reference resolution in dialogue systems, such as Kennington and Schlangen (2015). However, unlike Kennington and Schlangen, who explicitly train an object recognizer associated with each word of interest, with at least 65 labeled positive training examples per word, our model does not have any comparable form of supervision and our data exhibits much lower frequencies of object and word (co-)occurrence. Moreover, reference resolution is only an aspect of what we do: Besides being able to associate a word with a visual extension, our model is simultaneously learning word representations that allow us to deal with a variety of other tasks—for example, as mentioned above, guessing the appearance of the object denoted by a new word from a purely verbal description, grouping concepts into categories by their similarity, or having both abstract and concrete words represented in the same space.

6 Conclusion

Our very encouraging results suggest that multimodal distributed models are well-suited to simulating human word learning. We think the most pressing issue to move ahead in this direction is to construct larger corpora recording the linguistic and visual environment in which children acquire language, in line with the efforts of the Human Speechome Project (Roy, 2009; Roy et al., 2015). Having access to such data will enable us to design agents that acquire semantic knowledge by leveraging all available cues present in multimodal communicative setups, such as learning agents that can automatically predict eye-gaze (Recasens* et al., 2015) and incorporate this knowledge into the semantic learning process.

Acknowledgments

We thank Marco Marelli for useful advice and Brent Kievit-Kylar for help implementing the Best-F measure. We acknowledge the European Network on Integrating Vision and Language for a Short-Term Scientific Mission grant, awarded to Raquel Fernández to visit the University of Trento.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR Conference Track*, San Diego, CA. Published

online: http://www.iclr.cc/doku.php?id=iclr2015:main.

Marco Baroni, Alessandro Lenci, and Luca Onnis. 2007. ISA meets Lara: An incremental word space model for cognitively plausible simulations of semantic learning. In *Proceedings of the ACL Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 49–56.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Jia Deng, Wei Dong, Richard Socher, Lia-Ji Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of CVPR*, pages 248–255, Miami Beach, FL.

Afsaneh Fazly, Afra Alishahi, and Suzanne Stevenson. 2010. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34:1017–1063.

Michael Frank, Noah Goodman, and Joshua Tenenbaum. 2007. A Bayesian framework for cross-situational word-learning. In *Proceedings of NIPS*, pages 457–464, Vancouver, Canada.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Ákos Kádár, Afra Alishahi, and Grzegorz Chrupała. 2015. Learning word meanings from images of natural scenes. *Traitement Automatique des Langues*. In press, preprint available at http://grzegorz.chrupala.me/papers/tal-2015.pdf.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the Conference for the Association for Computational Linguistics (ACL)*.

Brent Kievit-Kylar and Michael Jones. 2011. The Semantic Pictionary project. In *Proceedings of CogSci*, pages 2229–2234, Austin, TX.

Brent Kievit-Kylar, George Kachergis, and Michael Jones. 2013. Naturalistic word-concept pair learning with semantic spaces. In *Proceedings of CogSci*, pages 2716–2721, Berlin, Germany.

Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. In *Proceedings of the NIPS Deep Learning and Representation Learning Workshop*, Montreal, Canada. Published online: http://www.dlworkshop.org/accepted-papers.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of NIPS*, pages 1097–1105, Lake Tahoe, Nevada.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of NAACL*, pages 153–163, Denver, CO.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk*. Lawrence Erlbaum Associates, 3rd edition.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. http://arxiv.org/abs/1301.3781/.

Adria Recasens*, Aditya Khosla*, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*. * indicates equal contribution.

Leslie Rescorla. 1980. Overextension in early language development. *Journal of Child Language*, 7(2):321–335.

Brandon C. Roy, Michael C. Frank, Philip DeCamp, Matthew Miller, and Deb Roy. 2015. Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112(41):12663–12668.

Deb Roy. 2000. A computational model of word learning from multimodal sensory input. In *Proceedings of the International Conference of Cognitive Modeling (ICCM2000), Groningen, Netherlands*.

Deb Roy. 2009. New horizons in the study of child language acquisition. In *Proceedings of Interspeech*.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of ACL*, pages 721–732, Baltimore, Maryland.

Tanya Stivers and Jack Sidnell. 2005. Introduction: Multimodal interaction. *Semiotica*, pages 1–20.

John Trueswell, Tamara Medina, Alon Hafri, and Lila Gleitman. 2013. Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, 66(1):126–156.

Chen Yu and Dana H. Ballard. 2007. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165.

Chen Yu and Linda B. Smith. 2013. Joint attention without gaze following: human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PLoS ONE*, 8(11).

C. Yu. 2005. The emergence of links between lexical acquisition and object categorization: A computational study. *Connection Science*, 17(3):381–397.