COPING WITH COMBINATORIAL UNCERTAINTY IN WORD LEARNING: A FLEXIBLE USAGE-BASED MODEL

PIETER WELLENS

VUB AI-Lab, Pleinlaan 2,
1050 Brussels, Belgium
pieter@arti.vub.ac.be

Scaling up the complexity of a language game brings about a towering scale-up in the uncertainty the agents are faced with when acquiring (lexical) form-meaning associations. The two most prominent assumptions influencing the uncertainty in models on word meaning concern (1) meaning transfer and (2) whether a form can be associated with only one part of meaning or any subset of parts of meaning. If meaning has internal structure (e.g. sets of attributes) this second assumption amounts to whether a form can be associated with only one attribute, giving rise to linear uncertainty, or any subset of attributes, resulting in exponential uncertainty. We first present a short overview of different models that each tried to tackle at least one of these assumptions. We propose a new model borrowing ideas from many of these models that can handle the exponential increase in uncertainty when removing both assumptions and allows scaling towards very large meaning spaces (i.e. worlds).

1. Introduction and overview of previous models

Language learners are faced with immense uncertainty when trying to acquire lexical form meaning associations. Taking into account the enormous diversity found in human natural languages (Haspelmath, Dryer, Gil, & Comrie, 2005) or the subtleties in word use as described in (Fillmore, 1977) it becomes clear that few a priori assumptions (biases) can be made by the learner.

Two assumptions play a pivotal role in trying to capture the amount of uncertainty in acquiring form-meaning associations in a model. The first is whether meaning is explicitly transfered from speaker to hearer to allow near perfect one-shot learning and alignment. These models cannot say much, if anything at all, about word meaning but focus primarily on conventionalization dynamics. The acquisition or bootstrapping task can be regarded as a mapping problem, mapping form to meaning as in Siskind (1996). The second assumption can now be formulated as to whether such a mapping is restricted to one-to-one, essentially giving the set of word meanings in advance or whether word meaning has to be shaped from a set of more atomic parts of meaning, allowing a form to map to any subset of those atomic parts of meaning as depicted in 1. In a one-to-one mapping model the amount of uncertainty when confronted with a new form scales linearly with

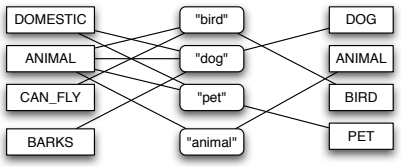

Figure 1. In the middle is a list of four forms. The right-hand side depicts how these would be associated in a model that assumes a one-to-one mapping of form and meaning. On the left-hand side you see the associations when this assumption is removed. Note that the atoms of meaning are different.

the number of attributes[a], when removing the one-to-one assumption the scale-up becomes exponential.

De Beule et al. (2006), Vogt and Divina (2007) and Siskind (1996) all have proposed models based on a cross-situational learning approach to bootstrap a shared lexicon without explicit meaning transfer. None of these models removed the one-to-one mapping assumption allowing only a linear scale in uncertainty. Other models such as (Smith, 2005) and (Steels, Kaplan, McIntyre, & Van Looveren, 2001) allow the attributes to be continuous (called channels) but a form usually only maps to one interval on one channel. Due to the continuous attributes the hypothesis space is infinite but in practice the uncertainty is heavily reduced by the one form to one interval assumption combined with few channels and few objects. As an example, when removing the one-to-one mapping assumption and meaning represented as a set of boolean attributes, a novel form uttered in combination with a topic containing 60 unexpressed attributes, can denote any of all subsets of these 60 attributes resulting in approximately 1.152921×10^{18} possible word meanings. It becomes worse as novel words are most often heard in multi word utterances and words known by both speaker and hearer do not necessarily have equal meaning (in some cases lexical coherence can be quite low) and in robotic experiments attributes are never exactly equal due to noise and different perspectives. In contrast to this massive amount of uncertainty, in De Beule et al. (2006) the maximum amount of hypotheses for a form-meaning association is 100 and at the first exposure it would only be 5. In Vogt and Divina (2007) the worst case scenario at first exposure results in 26 hypotheses.

Other models such as those presented by De Beule and Bergen (2006), Steels and Loetzsch (2007), Steels and Kaplan (2000) in different ways remove both assumptions but they tend to keep the hypothesis space small (i.e. the number of possibilities when hearing a new form). For example the experiments in (De Beule

[a]Since I review different models and they use different terminology I will call the atoms of meaning *attributes*. In De Beule, De Vylder, and Belpaeme (2006) these are called objects, in Vogt and Divina (2007) perceptual features and in Siskind (1996) they would be referred to as concepts

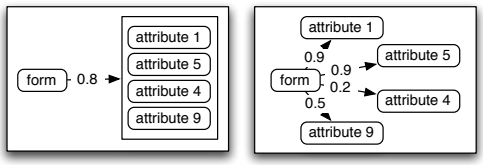

Figure 2. Left an association between form and meaning as in (De Beule & Bergen, 2006) scoring the complete subset. Right the refinement suggested in the proposed model.

& Bergen, 2006) consist of only 60 topics in total and only 10 attributes (there called predicates). It must be said that these papers did not address scale-up and therefore do not claim to handle it.

Wellens and Loetzsch (2007) present a model that puts uncertainty at the core of language representation, processing, alignment and learning. In the current paper we do not go into the implementation details since they have been detailed in the paper above. We wish only to present this model as an elegant solution when removing both assumptions, capable of handling exponential uncertainty in very large meaning spaces. The model is most heavily influenced by psycho-linguistics (Tomasello, 2003), cognitive linguistics (Langacker, 2002; Croft & Cruse, 2004) and other computer models (Batali, 1998; Steels & Belpaeme, 2005; De Beule & Bergen, 2006; De Beule et al., 2006) and is therefore heavily indebted to all these works.

2. Overview of the model

The agents engage in series of guessing games (Steels et al., 2001). Every game consists of two agents, a randomly assigned speaker and hearer, that share a joint attentional frame (the context). The goal of the speaker is to draw the hearer's attention to a randomly chosen object (the topic) using only language. The hearer, after interpretation, points to which he believes the speaker picked as the topic and if incorrect the speaker points to the topic. Meaning is modeled as a set of boolean attributes. As explained in the introduction pointing by the speaker is never explicit meaning transfer (i.e. transferring the intended set of attributes) but just to the object as a whole leaving the hearer with an exponential amount of referential uncertainty since the potential meaning could be any subset of attributes of the pointed object.

With the number of hypotheses per form well over the billions it is clear an agent cannot enumerate these possibilities and score them separately neither can he make series of one-shot guesses and hope for the best since finding the correct meaning would be like winning in lottery. As is clearly demonstrated in (Tomasello, 2003) children's word learning does not resemble any of these two

approaches as children seem to gradually home in on the correct meaning[b]. It is thus clear the agents require a continuously refining representation of the hypothesized meaning.

A first step is to separately score every attribute in a form-meaning association instead of keeping only one scored link per word as in (De Beule & Bergen, 2006) (see figure 2). This allows alignment, which is performed at the end of every game, to be more subtle. For every word in the utterance the scores of all associated attributes get updated. Those attributes that co-occured in topic and the used words get incremented (entrenchment) and the others get decremented (erosion) resembling the Langacker (2002) dynamic usage based model. When a score (confidence) hits zero the association gets pruned allowing word meaning to become more general. New associations can also be added but only by the hearer in failed games by adding all unexpressed attributes of the topic to all uttered words. It might also help to interpret the association score as denoting the confidence that the associated attribute is actually part of the meaning. In its most general form this sort of updating could be regarded as cross-situational learning but in other cross-situational models like (De Beule et al., 2006; Vogt & Divina, 2007; Siskind, 1996; Smith, 2005) the different associations are *competing* because of the one-to-one mapping assumption. It is crucial to understand that in the current model different attributes associated with the same form are *not* competing since the model allows a combinatorial mapping between form to attributes.

The key idea during production is that the internal structure of the word meaning allows for the calculation of similarity between meanings. Looking back at figure 1 an agent with the left lexicon 'knows' that a "dog" is also a "pet" since "pet" is associated with a subset of the attributes of dog. The similarity measure used in the model is an overlap metric additionally taking into account the entrenchment (the score) of the attributes and an extra weight that is also gradually learned by the agents. More details about the implementation can be found in Wellens and Loetzsch (2007). Using this similarity measure production amounts to measuring similarity between the topic and the word meanings choosing that combination of words that is most similar to the topic and least similar to the other objects in the context. This makes for context sensitive utterances and can be interpreted as an on-the-fly discrimination. The most important corollary of using a similarity measure is the great flexibility in word combination, especially in the beginning when the attributes are not very entrenched. This flexibility allows the agents to use (combinations of) words that do not fully conform to the meaning to be expressed resembling what Langacker (2002) calls *extension*.

If categorization is interpreted as the attempt to "recognize" a stan-

[b]Since in the experiments the agents need to bootstrap their lexicons there is no apriori 'correct' meaning but they need to shape the best concepts while bootstrapping their communication system as well.

> dard S in a target T, then instantiation represents the privileged case where this happens unproblematically, and extension constitutes recognition accomplished only with a certain amount of "strain". The source of the strain is that, for the S to be recognized in a target which does not fully conform its specifications, the conflicting features of S somehow have to be suppressed or abstracted away from.

Combining the similarity based flexibility with the entrenchment and erosion effects during alignment the word meanings will gradually be shaped in such a way to better conform future use requiring less "strain". When repeated over thousands of language games the word meanings will therefore gradually refine and shift until they capture frequently occurring structures (clusters) in the world effectively implementing a search through the enormous hypothesis space. Interpretation amounts to looking up all uttered words, combining their attributes and measuring similarity between the interpretation and the objects in the context. The hearer then points to the object with highest similarity score, again making interpretation very flexible.

Word invention is triggered if the best combination of words (the one most similar to the topic) is still more similar to another object in the context implying that the speaker himself would not even point to the topic he has chosen on interpreting his own sentence. The ability to interpret one's own sentence to diagnose potential misinterpretation turns out to be crucial in many models (Batali, 1998; Steels, 2003; Steels & Wellens, 2006). The speaker starts from the best utterance invents a new form and associates, with very low initial score, all unexpressed attributes of the topic with this new form. Even in this case it might not always be necessary to invent a new word because, due to erosion and entrenchment, word meanings can shift. Chances are that the words need some more time to be shaped into a more optimal meaning. To incorporate this the chance a new word is invented is proportional to the similarity of the best found combination of words to the topic. The closer the combination is to the topic, the less likely a new word will be invented because the more likely erosion and entrenchment might take care of it. The hearer, when adopting novel words, first interprets all known words and associates all unexpressed attributes with all novel forms.

3. Experiments and observations

Figure 3 shows the result of two experimental runs, both using a population of 20 agents, the difference being that the left features a small world consisting of only 32 objects each composed of 10 boolean attributes with context sizes between 4 and 10 objects and the right involves a world containing 8124 distinct objects, each described by 100 boolean attributes and context sizes between 5 and 20. The main difference is not the total amount of objects but the scale-up from 10 to 100 attributes scaling from a hypothesis space of size 1024 to one of size 2^{100} or 1.26765×10^{30}. The following measures are depicted:

Communicative Success (left axis): A running average (window of 500) of communicative success as measured by the agents. A game is considered successful if the hearer points to the correct topic. It is therefore different from communicative accuracy as employed in (Vogt & Divina, 2007; Siskind, 1996).

Lexicon Size (right axis): Represents the average number of words in the lexicon of the agents.

Lexicon Coherence (left axis): Measures the similarity (using the same similarity measure the agents use) between the lexicons of the agents. Coherence of 1 indicates that for all words all agents have the exact same attributes associated. In many cases it is (and makes sense to be) lower than one since it is not required to have the exact same meanings, although of course the higher it is the more robust they are in changing environments. Note that this is a measure the agents cannot access themselves since they cannot look into each others brains.

Experiments on the small meaning space are averaged over 5 runs, the ones on the larger meaning space are not due to the massive amount of processing power required for those experiment.

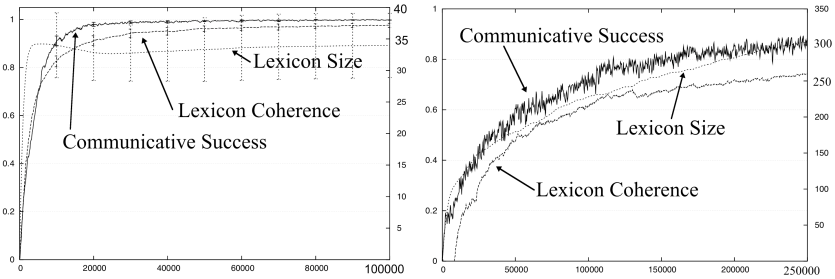

Figure 3. Left shows the performance of the proposed model on a small world, right for a much larger world. Clearly the model reaches high levels of communicative success and lexical coherence when scaling the uncertainty.

Although the second experiment obviously requires more games it is clear that in both experiments agents fairly quickly rise to high levels of communicative success, achieve high coherence and word invention stabilizes fairly quickly.

As a comparison we ran both experiments using the model described in (De Beule & Bergen, 2006) which differs from the proposed model in that it only scores the complete meaning and not the individual components of meaning as in figure 2. Results are shown in figure 4 and although the experiment with the small meaning space eventually reaches high success the agents in the scaled-up world

fail at bootstrapping a communication system. Also note that even in the small world the agents using this second approach only reach 20 % communicative success by game 20000 while with the proposed model they have already attained close to 99 % communicative success by then.

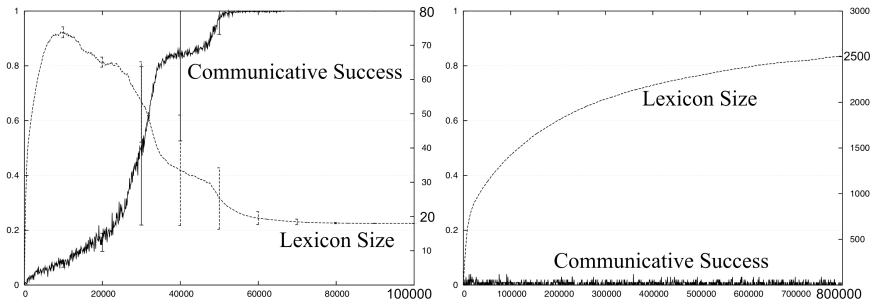

Figure 4. Left shows the performance of a model proposed in (De Beule & Bergen, 2006) on a small meaning space, right for a much larger space. The model achieves success on the small one, but fails to scale to the larger meaning space.

4. Conclusion

The model presented here has borrowed from many sources such as psycholinguistics and child language acquisition (Tomasello, 2003) cognitive linguistics (Langacker, 2002; Croft & Cruse, 2004) and other computer models (Batali, 1998; Steels & Belpaeme, 2005; De Beule & Bergen, 2006). It has shown that a population of agents can bootstrap a shared lexicon when scaling up the amount of uncertainty by removing two fundamental assumptions commonly found in models of lexicon formation. By scoring each component of meaning and not considering these components as competitors (because of a one-to-one mapping assumption) but rather use similarity based language processing and subtle alignment a population can effectively cope with immense amounts of uncertainty and bootstrap a shared lexicon in very large meaning spaces.

Acknowledgements

The research reported here has been conducted at the Artificial Intelligence Laboratory of the Vrije Universiteit Brussel (VUB). Pieter Wellens is funded by FWOAL328. I would like to thank Remi van Trijp, Joris Bleys, Joachim De Beule and Luc Steels for their useful commentaries.

References

Batali, J. (1998). Computational simulations of the emergence of grammar. In J. R. Hurford, M. S. Kennedy, & C. Knight (Eds.), *Approaches to the evo-*

lution of language: Social and cognitive bases. Cambridge: Cambridge University Press.

Croft, W., & Cruse, A. D. (2004). *Cognitive linguistics (cambridge textbooks in linguistics).* Cambridge: Cambridge University Press.

De Beule, J., & Bergen, B. K. (2006). On the emergence of compositionality. In *Proceedings of the 6th evolution of language conference* (p. 35-42).

De Beule, J., De Vylder, B., & Belpaeme, T. (2006). A cross-situational learning algorithm for damping homonymy in the guessing game. In L. M. R. et al. (Ed.), *Artificial life x* (p. 466-472). MIT Press.

Fillmore, C. J. (1977). Scenes-and-frames semantics. In A. Zampolli (Ed.), *Linguistic structures processing* (p. 55-81). Amsterdam: North-Holland.

Haspelmath, M., Dryer, M., Gil, D., & Comrie, B. (Eds.). (2005). *The world atlas of language structures.* Oxford: Oxford University Press.

Langacker, R. W. (2002). A dynamic usage-based model. In *Usage based models of language.* Stanford, California: CSLI Publications.

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39-91.

Smith, A. D. M. (2005). The inferential transmission of language. *Adaptive Behavior*, *13*(4), 311-324.

Steels, L. (2003). Language re-entrance and the inner voice. *Journal of Consciousness Studies*, *10*, 173-185.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*(4), 469-89. (Target Paper, discussion 489-529)

Steels, L., & Kaplan, F. (2000). Aibo's first words: The social learning of language and meaning. *Evolution of Communication*, *4*(1), 3–32.

Steels, L., Kaplan, F., McIntyre, A., & Van Looveren, J. (2001). Crucial factors in the origins of word-meaning. In A. Wray (Ed.), *The transition to language* (p. 252-271). Oxford University Press.

Steels, L., & Loetzsch, M. (2007). Perspective alignment in spatial language. In K. Coventry, T. Tenbrink, & J. Bateman (Eds.), *Spatial language and dialogue.* Oxford: Oxford University Press.

Steels, L., & Wellens, P. (2006). How grammar emerges to dampen combinatorial search in parsing. In P. Vogt (Ed.), *Eelc 2: Symbol grounding and beyond* (Vol. 4211, p. 76-88). Berlin Heidelberg: Springer Verlag.

Tomasello, M. (2003). *Constructing a language. a usage based theory of language acquisition.* Harvard University Press.

Vogt, P., & Divina, F. (2007). Social symbol grounding and language evolution. *Interaction Studies*, *8*(1).

Wellens, P., & Loetzsch, M. (2007). Flexible word meaning in embodied agents. In *Social learning in embodied agents.* Lisbon.