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

Agents in the process of bootstrapping a shared lexicon face immense uncertainty. The problem that an agent cannot point to meaning but only to objects, represents one of the core aspects of the problem. Even with a straightforward representation of meaning, such as a set of boolean features, the hypothesis space scales exponential in the number of primitive features. Furthermore, data suggests that human learners grasp aspects of many novel words after only a few exposures. We propose a model that can handle the exponential increase in uncertainty and allows scaling towards very large meaning spaces. The key novelty is that word learning or bootstrapping should not be viewed as a mapping task, in which a set of forms is to be mapped onto a set of (predefined) concepts. Instead we view word learning as a process in which the representation of meaning gradually shapes itself, while being usable in interpretation and production almost instantly.

1. Introduction

Word learning is commonly viewed as a mapping task, in which the learner has to map a set of forms onto a set of concepts (Bloom, 2000; Siskind, 1996). While mapping might seem more straightforward than having to shape word meanings, it is in fact more difficult and lies at the root of many problems.

The view that word learning corresponds to mapping forms onto concepts is commonly accompanied by claims that a learner is endowed with several biases (constraints) that guide him toward the right mapping (Markman, 1989). Whether these constraints are language specific is yet another debate (Bloom, 2001). While this approach recognises the uncertainty it largely circumvents it by invoking these constraints. Another possibility is to propose some form of cross situational learning where the learner enumerates all possible interpretations and prunes this set when new data arrives. This second approach would seem to have a problem explaining fast mapping, since it takes a large amount of time before the initial set of hypotheses can be pruned to such an extent that it becomes usable.

To be clear, we are not unsympathetic to the idea of word learning constraints, but we believe that it is only when viewing word learning as mapping that the constraints become as inescapable as they seem. In this publication we try to

show that by trading the mapping view for a more organic, flexible approach of word learning (in line with Bowerman and Choi (2001)), the constraints become less cardinal. Moreover, the enormous diversity found in human natural languages (Haspelmath, Dryer, Gil, & Comrie, 2005; Levinson, 2001) and the subtleties in word use (Fillmore, 1977) suggest that language learners can make few apriori assumptions and even if they would, they still face a towering uncertainty when homing in on more subtle aspects of word meaning and use.

Some developmental psychologists emphasize human proficiency in interpreting the intentions of others (Tomasello, 2003) or our endowment with a theory of mind (Bloom, 2000). While being supportive of these ideas and even taking some for granted in our experimental set-up, it is important to understand that intention reading is no telepathy. It might scale down the problem, but not entirely solve it. Any of these skills have to be accompanied by a model capable of coping with immense uncertainty in large hypothesis spaces.

Siskind (1996) and others propose models based on cross situational learning to bootstrap a shared lexicon. Unlike the current experimental setup their experiments do not address an exponential scale-up in the number of hypotheses. Other models such as De Beule and Bergen (2006), Steels and Loetzsch (2007), Steels and Kaplan (2000) in different ways allow exponential scaling but tend to keep the hypothesis space small. For example the experiments in De Beule and Bergen (2006) are limited to 60 objects represented by 10 distinct features (there called predicates). These papers, however, do not address scale-up and therefore do not claim to handle it.

2. Overview of the model

Agents engage in series of guessing games (Steels, 2001). A guessing game is played by two agents, a randomly assigned speaker and hearer, sharing a joint attentional frame (the context). The speaker has to draw the hearer's attention to a randomly chosen object (the topic) using one or more words in its lexicon. After interpretation, the hearer points to which he believes the speaker intended. In case of failure, the speaker corrects the hearer by pointing to the topic.

To investigate referential uncertainty, which is the problem that an agent cannot point to meaning but only to objects, we must ensure that multiple equally valid interpretations exist upon hearing a novel word. It follows that explicit meaning transfer (i.e. telepathy) or a non structured representation of meaning are to be avoided. Even with an elementary representation of meaning such as sets of primitive features the number of possible interpretations scales exponential in the number of features, given that word meaning can be any subset of these features[a]. For example, upon hearing a novel word, sharing joint attention to an

[a] We do not claim such a representation to be realistic, but we believe it is the minimal requirement that suits our current needs for investigating the problem of referential uncertainty.

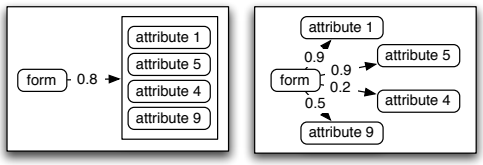

Figure 1. Left an association between form and meaning as in common in many models of lexicon formation, scoring the complete subset. Right the refinement suggested in the proposed model, which is related to fuzzy sets and prototype theory.

object represented by 60 boolean features, and having no constraints to favor particular interpretations the intended meaning could be any of $2^{60} = 1.153 \times 10^{18}$ possibilities. Confronted with numbers of such magnitude one wonders how a learner, given a stable input language, ever achieves in finding out the intended meaning, let alone a population of agents bootstrapping, from scratch, a shared lexical language. Word learning constraints seem to be the only viable way out.

With the number of hypotheses per novel word well over the billions a learner 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.

The first step towards a solution is to include uncertainty in the representation of word meaning itself. This is done by keeping an *(un)certainty score* for every feature in a form-meaning association instead of keeping only one scored link per word as in for example (De Beule & Bergen, 2006) (see figure 1). This representation is strongly related to both fuzzy set theory (Zadeh, 1965) and prototype theory (Rosch, 1973). A crucial difference with traditional cross situational learning approaches is that this representation avoids the need to explicitly enumerate competing hypotheses.

The key idea during language use is that a weighted similarity can be calculated between such representations. In the model we use a weighted overlap metric using the certainty scores as weights. In short, shared features increase similarity and the disjunct parts decrease it. Employing this similarity measure, production amounts to finding that combination of words of which the meaning is most similar to the topic and least similar to the other objects in the context. This results in context sensitive multi-word utterances and involves an implicit on-the-fly discrimination using the lexicon.

The most important corollary of using a similarity measure is the great flexibility in word combination, especially in the beginning when the features have low certainty scores. Thanks to this flexibility the agents can use (combinations of) words that do not fully conform the meaning to be expressed, resembling what Langacker (2002) calls *extension*. The ability to use linguistic items beyond their

specification is a necessity in high dimensional spaces to maintain a balance between lexicon size and coverage (expressiveness).

Interpretation amounts to looking up the meaning of all uttered words, taking the fuzzy union of their features and measuring similarity between this set and every object in the context. The hearer then points to the object with highest similarity, again making interpretation flexible.

Flexible use of words entails that in a usage event some parts of the meanings are beneficial and others are not. If all features of the used meanings are beneficial in expressing the topic it would not be extension but instantiation, which is rather the exception than the rule. As Langacker (2002) puts it, extension entails "strain" in the use of the linguistic items which in turn affects the meanings of the used linguistic items. This is operationalised by slightly shifting the certainty scores every time a word is used in production or interpretation. The certainty score of the features that raised the similarity are incremented and the others are decremented resembling the psychological phenomena of entrenchment and its counterpart erosion. Features with a certainty score equal or less than 0 are removed, resulting in a more general word meaning. In failed games the hearer adds all unexpressed features of the topic to all uttered words, thus making the meanings of those words more specific.

Combining similarity based flexibility with entrenchment and erosion, word meanings gradually shape themselves to better conform future use. Repeated over thousands of language games the word meanings progressively refine and shift, capturing frequently co-occurring features (clusters) in the world, thus effectively implementing a search through the enormous hypothesis space, capturing what is functionally relevant.

Word invention is triggered when the speaker's best utterance cannot discriminate the chosen topic. To diagnose possible misinterpretation the speaker interprets his own utterance before actually uttering it, which is crucial in many models (Batali, 1998; Steels, 2003). Given that his lexicon is not expressive enough, the speaker invents a new form (a random string) and associates to it, with very low initial certainty score, all so far unexpressed features of the topic. Because word meanings can shift, it might not be necessary to introduce a new word. Chances are that the lexicon needs a bit more time to be shaped further. Therefore the more similar the meaning of the utterance is to the topic, the less likely a new word will be introduced.

The hearer, when adopting novel words, first interprets all known words and associates, again with very low certainty scores, all unexpressed features with all novel forms.

3. Experimental results

In the multi-agent experimental setup we use a population of 25 agents endowed with the capacities described in the previous section. Machine learning data-sets

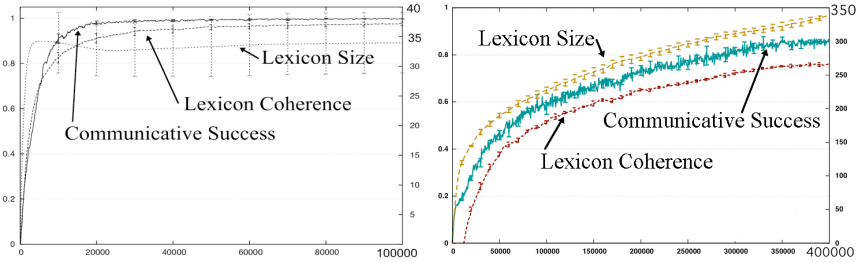

Figure 2. Left shows the performance of the proposed model on a small world (averaged over 5 runs), right for the much larger world (averaged over 3 runs) . Although the number of hypotheses scales exponential the agents attain high levels of communicative success and lexicon coherence while keeping reasonable lexicon size.

are used to obtain the large meaning spaces required to verify the claim that the model can scale to large hypothesis spaces. We use both a small data-set containing only 32 objects represented by 10 boolean features with context sizes between 4 and 10 objects, and a much larger data-set comprising 8124 objects represented by a total of 100 distinct boolean features and context sizes between 5 and 20 objects (Asuncion & Newman, 2007). This larger data-set confronts the agents with incredible amounts of uncertainty but the the results (figure 2) show that the model can manage this. 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 in Vogt and Divina (2007), Siskind (1996).

Lexicon Size (right axis): Represents the average number of words in the lexicons 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 features associated. It makes sense to be lower than 1 since it is not required to have the exact same meanings to be able to successfully communicate. The agents will not be aware of their (slightly) different meanings until a particular usage event confronts them with it.

As a comparison we ran a model that does not score the individual features, but instead keeps a score for the meaning as a whole as in figure 1 (left). It does not employ a similarity measure and updates scores based on communicative success instead of the more subtle entrenchment and erosion effects. Results show (figure 3) that the population can bootstrap a shared lexicon for small meaning spaces but

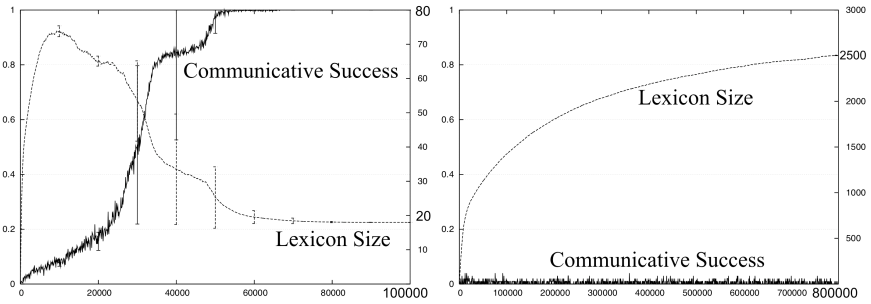

Figure 3. Both graphs show the performance of a model that doesn't score the individual features and does not use a similarity measure. Left for the small meaning space, right for the larger space. The model achieves success on the small one, but fails to scale to the larger meaning space.

cannot handle the scale up to the larger world. Also note that even in the small world the agents using this second model reach only 20% communicative success by game 20000 while with the proposed model they have already attained close to 99% communicative success by then.

Data from developmental psychology suggests that human learners can infer aspects of the meaning of a novel word after only a few exposures. The graphs in figure 2 do not give us any insight on these issues as they show the average of a population in the process of bootstrapping a lexicon. By adding a new agent to a population that has already conventionalised a shared lexicon we are able to shed light on the behaviour of the proposed model regarding this issue. We use the large world (8124 objects, 100 features), a stabilised population with an average lexicon size of some 100 words and measure for a newly added agent the average success in interpretation in relation to the number of exposures to the word (see figure 4). The graph shows the average success in interpretation (i.e. the new agent pointed correctly) of all words, in relation to the number of exposures. Due to the way success is measured the first exposure is always a failure and so average success is zero. Quite perplexing, on the second exposure a whopping 64% of the novel words are used in a successful interpretation. Further exposures gradually improve this result and by the tenth exposure 70% of the words result in a successful interpretation. This is the more baffling taking into account that the other members of the population are unaware they are talking to a new agent, and thus use multi-word utterances, including difficult to grasp words.

4. Conclusion

The proposed model tries to capture and bring together some insights from cognitive linguistics (Langacker, 2002) and other computational models (Batali, 1998; Steels & Belpaeme, 2005; De Beule & Bergen, 2006), while taking for granted insights from developmental psychology (Tomasello, 2003) and criticising assump-

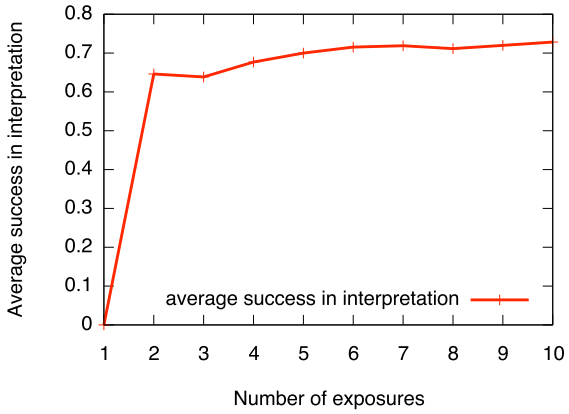

Figure 4. The graph shows the performance in interpretation of one new agent added to a stabilised population. Quite perplexing the average success in interpretation at the second exposure to a novel word is already 64%.

tions made by others (Bloom, 2000; Markman, 1989). The main strength of modelling is that it can operationalise ideas and so our main goal is in showing that a more organic view on word learning combined with flexible language representation, use and alignment results in a powerful idea, both for scaling to very large hypothesis spaces and arriving at operational interpretations after very few exposures. Although our model can be interpreted as Whorfian this is only so if you assume that word meanings and concepts are one and the same. We did not make this assumption and do not take a position regarding the relation of concepts and word meanings.

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 my supervisor Luc Steels and the referees for their useful comments.

References

Asuncion, A., & Newman, D. (2007). *UCI machine learning repository.*

Batali, J. (1998). Computational simulations of the emergence of grammar. In J. R. Hurford, M. S. Kennedy, & C. Knight (Eds.), *Approaches to the evolution of language: Social and cognitive bases.* Cambridge: Cambridge University Press.

Bloom, P. (2000). *How children learn the meanings of words.* MIT Press.

Bloom, P. (2001). Roots of word learning. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 159–181). Cambridge: Cambridge University Press.

Bowerman, M., & Choi, S. (2001). Shaping meanings for language: Universal and language-specific in the acquisition of spatial semantic categories. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 132–158). 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).

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.

Levinson, S. C. (2001). Language and mind: Let's get the issues straight! In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (p. 25-46). Cambridge: Cambridge University Press.

Markman, E. (1989). *Categorization and naming in children: problems of induction.* Cambridge, MA: Bradford/MIT Press.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *7*, 573-605.

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

Steels, L. (2001). Grounding symbols through evolutionary language games. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 211–226). London: Springer Verlag.

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

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).

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 338–353.