ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1764

Relational Grounded Language Learning

Leonor Becerra-Bonache[1] and **Hendrik Blockeel**[2] and
María Galván[1] and **François Jacquenet**[1]

Abstract. In the past, research on learning language models mainly used syntactic information during the learning process but in recent years, researchers began to also use semantic information. This paper presents such an approach where the input of our learning algorithm is a dataset of pairs made up of sentences and the contexts in which they are produced. The system we present is based on inductive logic programming techniques that aim to learn a mapping between n-grams and a semantic representation of their associated meaning. Experiments have shown that we can learn such a mapping that made it possible later to generate relevant descriptions of images or learn the meaning of words without any linguistic resource.

1 INTRODUCTION

Learning language models has been a research challenge for many years now. Grammatical Inference [7] has focused on that topic for more than 50 years. One of the main objectives of that research domain is to discover a grammar (or an associated automaton) that is a model of a set of positive (and sometimes negative) examples of sequences of words over an alphabet. Nevertheless, the learning process mainly aims to learn a syntactic structure without the help of any additional semantic information.

Since the early 2000s, methods for grounded language learning and semantic parser construction have been proposed [3, 4, 10, 11, 12]. These learn from pairs (S,M), where S is a sentence and M a meaning, a function that maps (parts of) S onto (parts of) M. In this setting, the training set must explicitly contain in M the meaning of (each part of) S; the learning cannot construct meanings by combining elements of M.

To overcome this problem some work has tried to directly learn from pairs (S,C) where S is a sentence and C is a representation of the context in which S is produced. [1] has been a first attempt to implement such an approach. Then, more recently, we presented in [2] a more efficient approach where a context is represented in first order logic and Inductive Logic Programming techniques are used to learn a mapping between n-grams (sequences of words) and their associated meanings. The main improvement with respect to previous approaches in grounded language learning is that the meanings are not provided in the training set, our learning algorithm is able to discover it by itself.

The system we present in this paper aims to improve [2] that was a proof of concept and we now want to study how such a model can deal with more realistic contexts, in noisy environments, and to observe various linguistic knowledge that can be discovered by such an approach. To make our system more realistic we decided to provide contexts in the form of images. That makes the construction of the dataset easier as the users do not have to manually provide the contexts which would otherwise be a laborious task. Thus our system learns from pairs (S,I) where S is a sentence that talks about a part of an image I.

In that way our system can be related to some deep learning approaches to the image captioning task as presented for example in [5, 6, 8, 9, 15] where their training sets are similar to ours as they are made up of pairs (sentence,image). Nevertheless, these approaches aims to learn a function that can map basic image features to ordered sequences of words. At the moment these approaches do not learn any meaning representation from the training set and it is not possible to use the learned model to do any kind of inference or reasoning. In our approach we use information about the objects of the image and build a first order logic representation of the meaning of n-grams. Thus we are able later to do some inference on that representation and it would be even possible to add a reasoning component.

We may also notice the work from Mateos Ortiz et al. [13] that uses some Machine Translation techniques to generate image description. In fact their model differs from ours in the sense that they need to build a parallel corpus (sentence,image) where each sentence has to be pre-processed by a Part-Of-Speech tagger. The main concern with such an approach is that linguistic resources are needed and we want to design an approach that does not need any linguistic resource, we want the learner to discover by itself the resources it needs.

2 OUR SYSTEM

As explained in section 1, the input of our system is a dataset made up of pairs (S,I) where S is a sentence related to a particular image I. As we use, in our experiments, the Abstract Scene Dataset provided by Zitnick et al. [16], we do not have to face computer vision problems such as object detection, semantic labelling, etc. which are known to be very hard problems. This dataset provides for each image, the set of all of its objects with some associated features. Thus, after a basic pre-preocessing step, each image I is transformed into a scene Sc. For representing scenes, we use a first-order logic based representation. Thus scenes are made up of a set of ground atoms. These atoms describe properties of, and relationships between, the objects of I. Thus, the input of the learner is a dataset made up of pairs (S,Sc) where S is a sentence related to a particular scene Sc.

The core algorithm of our system mainly aims to compute the meanings associated with all the n-grams of all the sentences of the whole dataset. We consider the meaning of an n-gram is whatever is

[1] Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France, email: {leonor.becerra,maria.galvan,francois.jacquenet}@univ-st-etienne.fr
[2] Department of Computer Science, KU Leuven, Belgium, email: hendrik.blockeel@cs.kuleuven.be

in common among all the situations in which it can be used. Thus, the meaning of an n-gram is a *context* that may be present or absent in a given scene. In our model, the learner constructs some generalized context represented by a set of atoms that may contain variables.

To compute the most specific generalization of a set of contexts $\{C_1, \ldots, C_n\}$ associated with a particular n-gram, we use the well known Plotkin's "least general generalization" (lgg) operator [14] usually used on first order clauses. Applied to contexts, the lgg has a simple mathematical description: given two contexts (set of atoms), it returns the *least general* context that *subsumes* both contexts. A context C_1 subsumes a context C_2 if and only if there exists a variable substitution that turns C_1 into a subset of C_2; C_1 is then also said to be more general than C_2. In fact, the lgg of two contexts returns all the properties that they have in common.

Nevertheless we can face a problem when, in a pair (S,Sc), there is an object that is referred in S but whose corresponding atom is not present in Sc. For example if the word "red" means red, then using it in a sentence associated with a context without red objects will cause the system to conclude that the color red is *not* common to all the contexts where "red" occurs, and therefore "red" cannot mean the color red. This leads to overgeneralization of meanings. To overcome this important problem **the first improvement** with respect to [2] has been the use of some heuristics. Instead of stating that the meaning of an n-gram G has to be common to *all* the scenes where this n-gram is used, without any exception, the algorithm repeatedly generalizes the meaning of G by computing its lgg with another randomly chosen scene until at least a certain percentage of the scenes are subsumed by that meaning.

A **second improvement** with respect to [2] is the way our system can learn the meaning of words without any linguistic resource. Indeed, in [2], a word was chosen to refer to a constant if and only if that constant was the only one remaining in the word's meaning, which was a basic strategy. Now, when the meaning of a word (a 1-gram) is updated, our program looks at the constants that occur in this meaning. Among these, it finds the constant whose occurrence "correlates" best with that of the word. If the "correlation" is high enough, this constant is then chosen to be the constant the word refers to.

Finally, **the third improvement** with respect to [2] concerns the way our system can generate all the relevant sentences associated with a given scene. To avoid generating trivial sentences that are true for most or all scenes and therefore not interesting, we defined a scoring function for sentences that takes into account the specificity of the n-grams that are chained together to form those sentences.

We made a series of experiments based on the *Abstract Scenes Dataset*, proposed by Zitnick et al. [16] that contains images of children playing outdoors and sentences that describe these images. The objective was to study the ability of our model to: (i) generate relevant sentences for a given scene and (ii) learn the meaning of words. In these experiments, we used a model learned from a corpus of 10.000 examples (in English). We first asked the system to generate all the sentences that were relevant for a given set of scenes. The sentences were obtained by chaining 5-grams. Choosing appropriate parameters to tune the core algorithm, we observed that, on average, 80% of the sentences generated were syntactically and semantically correct. Furthermore, as mention in the previous section, the sentences generated by our system try to be as specific as possible, that is, they do not state things that are true for almost all the scenes (and consequently, not interesting, such as "the sky is blue" or "the grass is green").

Concerning the meaning of words, by fine-tuning the correlation parameters of the system, we observed some improvements in terms of precision from 40% to 85% and in terms of recall from 8% to 25% with respect to [2].

3 CONCLUSION

In this paper we presented a grounded language learning system based on ILP techniques that can learn from datasets made up of pairs (S,C) where C is a set of ground atoms that represent the (partial) state of the world that is (partially) described by the sentence S. Our algorithm learns mappings between n-grams and most specific generalizations of the contexts common to the given n-grams in the dataset. We have shown that our system is able to learn the meaning of words without any linguistic resource. Having learned a language model from such a dataset, our system is able to use this model to generate relevant sentences that can describe some new scenes.

REFERENCES

[1] D. Angluin and L. Becerra-Bonache, 'Effects of meaning-preserving corrections on language learning', in *Proc. of the 15th Int. Conf. on Computational Natural Language Learning*, pp. 97–105, (2011).

[2] L. Becerra-Bonache, H. Blockeel, M. Galván, and F. Jacquenet, 'A first-order-logic based model for grounded language learning', in *Proc. of the 14th Int. Symposium on Advances in Intelligent Data Analysis*, pp. 49–60. LNCS 9385, Springer, (2015).

[3] D. L. Chen, J. Kim, and R. J. Mooney, 'Training a multilingual sportscaster: Using perceptual context to learn language', *Journal of Artificial Intelligence Research*, **37**, 397–435, (2010).

[4] D. L. Chen and R. J. Mooney, 'Learning to sportscast: a test of grounded language acquisition', in *Proc. of the 25th Int. Conf. on Machine Learning*, pp. 128–135. ACM, (2008).

[5] X. Chen and C. L. Zitnick, 'Mind's eye: A recurrent visual representation for image caption generation', in *Proc. of the Int. Conf. on Computer Vision and Pattern Recognition*, pp. 2422–2431. IEEE, (2015).

[6] K. Cho, A. C. Courville, and Y. Bengio, 'Describing multimedia content using attention-based encoder-decoder networks', *IEEE Transactions on Multimedia*, **17**(11), 1875–1886, (2015).

[7] C. de la Higuera, *Grammatical Inference, Learning Automata and Grammars*, Cambridge University Press, 2010.

[8] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig, 'From captions to visual concepts and back', in *Proc. of the Int. Conf. on Computer Vision and Pattern Recognition*, pp. 1473–1482, (2015).

[9] A. Karpathy, A. Joulin, and F-F. Li, 'Deep fragment embeddings for bidirectional image sentence mapping', in *Proc. of the Int. Conf. on Neural Information Processing Systems*, pp. 1889–1897, (2014).

[10] R. J. Kate and R. J. Mooney, 'Learning language semantics from ambiguous supervision', in *Proc. of the 22nd AAAI Conf. on Artificial Intelligence*, pp. 895–900. AAAI Press, (2007).

[11] J. Kim and R. J. Mooney, 'Generative alignment and semantic parsing for learning from ambiguous supervision', in *Proc. of the 23rd Int. Conf. on Computational Linguistics*, pp. 543–551, (2010).

[12] T. Kwiatkowski, S. Goldwater, L. S. Zettlemoyer, and M. Steedman, 'A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings', in *Proc. of the 13th Conf. of the European Chapter of the ACL*, pp. 234–244, (2012).

[13] L. G. Mateos Ortiz, C. Wolff, and M. Lapata, 'Learning to interpret and describe abstract scenes', in *Proc. of the Conf. of the North American Chapter of the ACL*, pp. 1505–1515, (2015).

[14] G. D. Plotkin, *Machine Intelligence 5*, chapter A note on inductive generalization, 153–163, Edinburgh University Press, 1970.

[15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, 'Show and tell: A neural image caption generator', in *Proc. of the Int. Conf. on Computer Vision and Pattern Recognition*, pp. 3156–3164. IEEE, (2015).

[16] C. L. Zitnick, R. Vedantam, and D. Parikh, 'Adopting abstract images for semantic scene understanding', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**(4), 627–638, (2016).