Learning Language Models from Images with ReGLL

Leonor Becerra-Bonache[1], Hendrik Blockeel[2], Maria Galván[1], and François Jacquenet[1](✉)

[1] Université de Lyon, UJM-Saint-Etienne, CNRS, Saint-Etienne, France
Francois.Jacquenet@univ-st-etienne.fr

[2] Department of Computer Science, KU Leuven, Leuven, Belgium

Abstract. In this demonstration, we present ReGLL, a system that is able to learn language models taking into account the perceptual context in which the sentences of the model are produced. Thus, ReGLL learns from pairs (Context, Sentence) where: Context is given in the form of an image whose objects have been identified, and Sentence gives a (partial) description of the image. ReGLL uses Inductive Logic Programming Techniques and learns some mappings between n-grams and first order representations of their meanings. The demonstration shows some applications of the language models learned, such as generating relevant sentences describing new images given by the user and translating some sentences from one language to another without the need of any parallel corpus.

1 Introduction

Learning language models has been a very active domain of research for a long time and Grammatical Inference, a subdomain of Machine Learning dedicated to that task, has produced a huge number of results in the literature. That has already led to the implementation of tools used in various applications (see [3] for an overview of the domain).

This research has mainly focused on learning language models from a syntactic point of view considering training sets only made up of strings of characters or sequences of words. Nevertheless, it seems obvious that human beings do not learn languages in that way. If we look at very young children starting to learn their native language, we can note that they are exposed to many sentences that refer to things in a perceptible scene. Thus, some work have been done to integrate some semantic information in the language learning process. The work from Chen et al. [2] is one example of this way of learning language models. Nevertheless, in this approach the meaning of each sentence has to be provided for each example of the training set. The ReGLL (Relational Grounded Language Learning) prototype proposes a different approach in which the meaning

H. Blockeel—Work supported by KU Leuven (sabbatical grant) and FWO-Vlaanderen (WOG "Declarative Methods in Informatics").

B. Berendt et al. (Eds.): ECML PKDD 2016, Part III, LNAI 9853, pp. 55–58, 2016.
DOI: 10.1007/978-3-319-46131-1_12

of each sentence is automatically discovered by the system, thanks to the context associated with it.

Some work has focused on learning to caption images using some deep learning approaches. The work from Karpathy et al. [4] is one example of such an approach. The main idea behind these approaches is to learn a function that ranks sequences of words given some images. It is different from the ReGLL prototype that is able to build a general semantic representation of the meaning of (part of) sentences. Doing in that way makes it possible later to use this representation to reason about the universe that has been described by the set of images and sentences of the training set.

2 The ReGLL prototype

Due to space limitations we cannot detail the theoretical and algorithmic aspects behind ReGLL, for that purpose, the reader may refer to [1].

The input of the system is a dataset D1 made up of pairs (I,S) where I is an image that has been built using a scene builder and S is a sentence that describes (part of) the image. The scene builder provides a set of cliparts and the user can drag and drop cliparts to design a new scene. A preprocessing step can then transform D1 in a dataset D2 made up of Prolog facts that provide pairs (C,L) where C is a set of grounded atoms that contains all the information about the objects of the image I and L is the list of words of the sentence S. During the learning step, the system takes the dataset D2 as an input and generates a language model. In the demo we provide three families of datasets where the sentences are written in English and Spanish. Thus the language models learned are subsets of language models for English and Spanish.

ReGLL is based on Inductive Logic Programming (ILP) techniques where each learning step uses the *least general generalization* (lgg) operator [5]. The basic idea behind the ReGLL engine is to traverse the training set and, given an n-gram NG ($1 \leq n \leq 8$) that appears in a sentence, generate a most specific generalization of all the contexts of NG in the training set. The process is iterated for each n-gram of the training set. During the learning process, the system learns the meaning of 1-grams (words) and then learns the meaning of n-grams ($n \geq 2$).

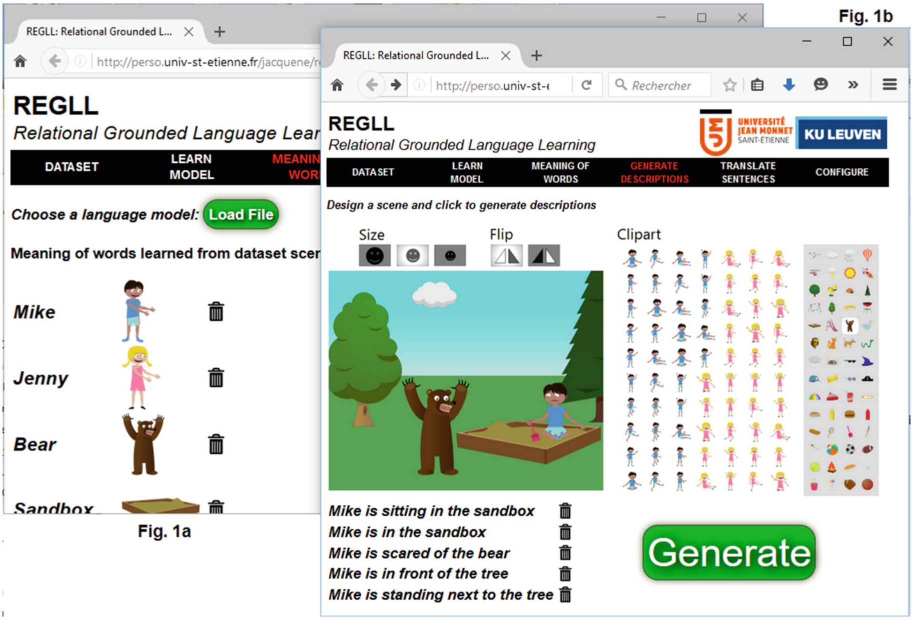
From an operational point of view, ReGLL is run through an interface that allows the user to act in various ways. One may: (i) load or design some training sets, (ii) learn some language models, (iii) load a language model to: (a) visualize the meaning of words, (b) generate relevant sentences given an image, (c) translate some sentences from a language L1 to a language L2 given the language models of L1 and L2.

3 Overview of the Demonstration

The demonstration is mainly based on the Abstract Scene Dataset built by Zitnick et al. [6] with a scene builder, nevertheless the attendees will be allowed to build some new datasets if they want to explore this functionality. It shows:

1. The way we can build a dataset from scratch using an abstract scene builder and then learn a language model from this dataset.
2. What can be done using a language model that has been learned.

Figure 1 shows two screenshots of the ReGLL system generating sentences describing images and showing the meaning of words it has discovered from a given dataset.



Choose a language model:

Meaning of words learned from dataset scer

Design a scene and click to generate descriptions

Mike is sitting in the sandbox
Mike is in the sandbox
Mike is scared of the bear
Mike is in front of the tree
Mike is standing next to the tree

Fig. 1. Screenshots of the ReGLL interface.

Three functionalities are thus mainly demonstrated:

- ***Visualizing the meaning of words.*** We can visualize what are the meanings that have been learned by ReGLL from a training set. Each word that has an associated meaning is displayed. The user may choose to delete some incorrect associations in order to help the system to be more efficient on the two other functionalities (see Fig. 1a).
- ***Describing images.*** Given an image that is built by the user using an image builder, the system may generate all the relevant sentences (ordered by decreasing relevance) that describe this image (see Fig. 1b).
- ***Translating sentences.*** We show that the language models learned by the system can be used to translate sentences written in a language L1 to sentences written in a language L2, while preserving the meaning. The languages available at the moment for this demonstration are English and Spanish.

Of course, the demonstration will allow attendees to look “inside the machine”. Indeed, it may be interesting for people familiar with Prolog to observe the code associated with the training sets, the main components of the learner and the language models.

We think this demonstration may be useful for people from both the academic and the industrial world working in the domain of natural language processing. For the academic audience, it may be interesting for people specialized in grammatical inference and people from the computation linguistic area. That can provide them some insights on how children learn from their environment. For the industrial audience, the core ideas behind ReGLL may be useful to design various tools. As the demonstration shows, such techniques can be used to design some tools able to generate descriptions of images, which can be very useful for blind people. In the domain of machine translation, our approach may be a new way to go beyond statistical machine translation that has well-known limitations that could be avoided by passing through a relational, more semantic representation. People from the domains of text summarization and Question-Answering may also find an interest in the ideas implemented in ReGLL.

4 Conclusion

The main goal of this demonstration of the ReGLL system is to prove the interest of learning language models not only from a syntactical point of view but also by taking advantages of semantic information related to the context in which the sentences of the language are produced. We expect attendees will actively use the system by themselves to explore its capabilities and discuss possible extensions that could integrate new functionalities they feel useful.

References

1. Becerra-Bonache, L., Blockeel, H., Galván, M., Jacquenet, F.: A first-order-logic based model for grounded language learning. In: Fromont, E., et al. (eds.) IDA 2015. LNCS, vol. 9385, pp. 49–60. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24465-5_5
2. Chen, D.L., Kim, J., Mooney, R.J.: Training a multilingual sportscaster: using perceptual context to learn language. J. Art. Int. Res. **37**, 397–435 (2010)
3. de la Higuera, C.: Grammatical Inference, Learning Automata and Grammars. Cambridge University Press (2010)
4. Karpathy, A., Joulin, A., Li, F.: Deep fragment embeddings for bidirectional image sentence mapping. In: Proceedings of the 28th NIPS Conference, pp. 1889–1897 (2014)
5. Plotkin, G.D.: A note on inductive generalization. In: Machine Intelligence 5, pp. 153–163. Edinburgh University Press (1970)
6. Zitnick, C.L., Vedantam, R., Parikh, D.: Adopting abstract images for semantic scene understanding. IEEE TPAMI **38**(4), 627–638 (2016)