ECAI 2010
H. Coelho et al. (Eds.)
IOS Press, 2010
© 2010 The authors and IOS Press. All rights reserved.
doi:10.3233/978-1-60750-606-5-1039

Analogical learning using dissimilarity between tree-structures

BEN HASSENA Anouar *and* MICLET Laurent[1]

Abstract. In Artificial Intelligence, analogy is used as a non exact reasoning technique to solve problems, for natural language processing, for learning classification rules, etc. This paper is interested in the analogical proportion, a simple form of the reasoning by analogy, and described its use in artificial learning of the syntactic tree (parsing) of a sentence.

1 Introduction

This paper belongs to several domains of artificial intelligence. Obviously, the first one is that of reasoning by analogy. Its application in artificial intelligence has been considered and tested early, among others in [3], [2], [8], [9].
The concept of analogy has been studied as one of the modality of reasoning since Aristotle ([6], [4]). It is also a reasoning by generalization, neither abductive nor inductive, which models a third form of learning. Recently, growing interest was manifested for a formal point of view on the analogy, the analogical proportion. This concept is rigorously defined and its application in representation spaces of various kinds has been developed with interest operational results. Its application to learning and the generation is conceptually simple, however, as in many areas of artificial intelligence, complexity of certain algorithmic problems remain to surmount.
In this paper we consider the problem of analogical learning using dissimilarity between trees, which we define as a multiple alignment of four ordered labeled trees, according to the notion of analogical proportion. We show how to extend the concept of alignment defined by [5] to aligning more than two trees. When four trees are considered, we propose to apply the concept of *analogical proportion* to trees and we extend it to that of *analogical dissimilarity*.
In the next section, we present the general notion of analogical proportion between four objects. We show our approach, starting from several original definitions to define analogical dissimilarity between trees. Section 3 gives two algorithms performing analogical tasks in the universe of trees. In the last section, we apply these algorithms to the learning of the syntactic tree (parsing) of a sentence.

2 Tree matching by analogical proportion

The analogical proportion is a relation between four objects which expresses that the way to transform the first object into the second is the same as the way to transform the third in the fourth. Let us call the objects O_1, O_2, O_3 and O_4. An analogical proportion is generally written by: "O_1 is to O_2 as O_3 is to O_4" and is denoted by $O_1 : O_2 :: O_3 : O_4$.

Definition 2.1 *An Analogical Proportion on a set $\mathbb{E}$ is a relation on $\mathbb{E}^4$ such that, for every 4-tuple A, B, C and D in relation in this order (which is denoted as $A : B :: C : D$), one has :*

1. $A : B :: C : D \Leftrightarrow C : D :: A : B$
2. $A : B :: C : D \Leftrightarrow A : C :: B : D$

An analogical equation *is a relation of the form $A : B :: C : X$, which has to be solved in X. It may have no, one or several solutions.*

When trees are considered, the question becomes, firstly, how to define an analogical proportion on trees and how to quantify the measure of similarity (actually, we use an *analogical dissimilarity*) in order to provide a more flexible method of learning by analogy.
With the same principles that [9] and [8] for strings, we show in this paragraph our methodology to define an analogical proportion between trees. The only prerequisite is that there exists an analogical proportion in the alphabet of the nodes labels augmented with the empty node label λ.

Definition 2.2 ***(Alignment between two trees)*** *An alignment between two trees T_1, T_2 whose labels are in Σ is a tree with labels in $(\Sigma) \times (\Sigma)/(\lambda, \lambda)$ which first projection is T_1, where the empty nodes λ are ignored and which second projection is T_2, where the empty node λ are ignored.*
Informally, an alignment represents a one to one node matching between two trees, in which some empty nodes may be inserted. The cost of an alignment is the sum of all nodes matching costs.

This definition can straightforwardly be extended to the alignment of any number of trees. When aligning four trees, we can apply the concept of analogical proportion to trees [1].

Definition 2.3 ***(Analogical proportion between trees)*** *Let x, y, z and t be four trees whose labels are in Σ. We suppose that an analogical proportion exists in Σ_λ. We say that these trees are in analogical proportion if there is an alignment of the four trees x', y', z' and t', with labels in Σ_λ^4, such that:*

- *For every node i of the alignment, the analogical proportion $x_i : y_i :: z_i : t_i$ of the labels holds true.*

Definition 2.4 ***(Tree analogical equation)*** *T_4 is a solution of the analogical equation $T_1 : T_2 :: T_3 : X$ if and only if the analogical proportion ($T_1 : T_2 :: T_3 : T_4$) holds true.*

We give now two originals algorithms to *(i)* validating proportions between trees and *(ii)* solving proportional equations on trees. In the latter case, as we will see in the following section, it is useful to define the concept of an *approximate solution* to an analogical equation.

[1] ENSSAT / IRISA, Lannion, France, email: {benhasse, miclet}@enssat.fr

2.1 Analogical Dissimilarity

In this section, we are interested in defining what could be a relaxed analogy, which linguistic expression would be "A is to B almost as C is to D". To remain coherent with our previous definitions, we measure the term "almost as" by some positive real value, equal to 0 when the analogical proportion is true, and increasing when the four objects are less likely to be in proportion. We call this value "analogical dissimilarity", in short *AD*.

Definition 2.5 ***(AD between trees)*** *Let X, Y, Z and T be four trees with labels $\in \Sigma_\lambda$. The analogical dissimilarity $AD(X,Y,Z,T)$ is the cost of the alignment of minimum cost between the four trees. This alignment is a tree A^4. We have: $AD(X,Y,Z,T) = \sum DA(x_i,y_i,z_i,t_i)$, with $i \in \left[1..\left|A^4\right|\right]$ and x_i, y_i, z_i and $t_i \in \Sigma_\lambda$.*
Coherence with analogy :
$DA(X,Y,Z,T) = 0 \Leftrightarrow X : Y :: Z : T$.
$DA(X,Y,Z,T) = DA(Z,T,X,Y) = DA(X,Z,Y,T)$.

Definition 2.6 ***(Approximate solution to an analogical equation)*** *Let $T_1 : T_2 :: T_3 : X$ be an analogical equation in trees. The set of best approximate solution to this equation is given by :*

$$X = \{\, x \;:\; x = ArgMin\ AD(T_1,T_2,T_3,x)\}$$

3 Algorithms on trees

3.1 AnaTree algorithm

This section is devoted to study an implementation of our definition 2.5. The algorithm we propose is based on dynamic programming. It progresses in synchrony in the four trees to build an optimal analogical alignment. The input of this algorithm will be the alphabet Σ' in which is defined an Analogical Dissimilarity on a 4-tuple of labels. The output is the AD between four trees. The algorithm covers fifteen possible quadruples alignments of forests to produce how much (at least) four trees "miss" the analogical proportion. This algorithm runs in a time complexity $O(|T|^4 \times (\text{ degree } (T))^4)$ its correctness is demonstrated by recurrence, as a generalization of Jiang algorithm [5] to align two trees.

3.2 SolvTree algorithm

When one of the four elements is unknown, the analogical proportion transforms into an analogical equation. For example, to solve the equation on letters $a : b :: a : x$, we need to produce all the letters x satisfying the analogical proportion, which is here reduced to $x = b$. When there is no solution, the notion of analogical dissimilarity will allow to discover an approximate solution. While this resolution of an analogical equation is trivial between letters, it is not straightforward to design an algorithm able to solve this kind of equation on trees, in particular when looking for an approximate solution if necessary. The algorithm we propose produces all the best exact or approximate solutions. The principle is to achieve an alignment of the three trees by browsing in each step eight possible cases. In each case, we calculate the cost of predicting the node label of the tree to be generated. Actually, we save at each step not only the cost of prediction, as shown above, but also the node(s) label(s) x found by analogical resolution along the optimal way of progression. When the calculation is finished, a backtracking gives us the optimal generated tree with a minimum analogical dissimilarity. The complexity of this algorithm is $O(|T|^3 \times (\text{ degree } (T))^3)$ in time.

4 Application to analogical syntactic parser

We consider a sentence P_0, which sequence S_0 of grammatical categories is known and which syntactic structure T_0 is searched for. Let AP a learning set of sentences (S,T), each sentence consisting of a sequence and a syntactic structure. The process of prediction by analogy of the parse tree T_0 is as follows:
1. Search for a triple of sentences (P_1, P_2, P_3) with sequences (S_1, S_2, S_3) and syntactic structures (T_1, T_2, T_3) such as the sequences S_0, S_1, S_2 and S_3 define an exact analogical proportion.

$$S_0 : S_1 : S_2 : S_3$$

2. Our hypothesis assumes that if the sequences are in analogy, so are the structures. Hence, we predict T_0 from the resolution of the analogical equation on trees: $x : T_1 :: T_2 : T_3$.
The corpus at our disposal consists of 316 sentences extracted from the base The Wall Street Journal Penn Treebank [7]. When the data available are limited, as it is the case here, the cross-validation technique can be used. Preliminaries results give an exact or almost exact restitution of the parsing tree from the sequence in 82 % of cases.

5 Conclusion

In this paper, we have proposed a new matching approach between trees using analogical proportion. We have extended to trees the concept of analogical dissimilarity, which measures the cost of matching when the trees are not in exact analogy. Two algorithms have been implemented: the first measures the AD between four trees, the second builds from three trees the fourth tree at minimum AD. We have considered as a first evaluation an experiment in automatic parsing. Then, the results with this evaluation protocol seem encouraging. We do not want to be competitive with other parsing systems but rather to show that a consistent use of analogy in learning can produce better results than other lazy learning approaches, like the nearest neighbor method.

REFERENCES

[1] A. Ben Hassena and L. Miclet, 'Dissimilarité analogique et apprentissage d'arbres', in *CAp'09: 11ème Conférence d'apprentissage CAp 2009*, pp. 121–132, (2009).

[2] J. G. Carbonell, 'Learning by analogy: Formulating and generalizing plans from past experience', in *Machine Learning: An Artificial Intelligence Approach*, eds., R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, 137–161, Springer, Berlin, Heidelberg, (1984).

[3] T. Evans, 'A heuristic program to solve geometry analogy problems', in *Semantic Information Processing*, MIT Press, Cambridge, (1968).

[4] K. Holyoak, 'Analogy', in *The Cambridge Handbook of Thinking and Reasoning*, chapter 6, Cambridge University Press, (2005).

[5] Tao Jiang, Lusheng Wang, and Kaizhong Zhang, 'Alignment of trees - an alternative to tree edit', in *CPM '94: Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching*, pp. 75–86, London, UK, (1994). Springer-Verlag.

[6] Y. Lepage, *De l'analogie rendant compte de la commutation en linguistique*, Grenoble, 2003. Habilitation à diriger les recherches.

[7] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini, 'Building a large annotated corpus of english: the penn treebank', *Comput. Linguist.*, **19**(2), 313–330, (1993).

[8] L. Miclet, S. Bayoudh, and A. Delhay, 'Analogical dissimilarity: Definition, algorithms and two experiments in machine learning', *journal of Artificial Intelligence Research*, **32**, 793–824, (2008).

[9] N. Stroppa and F. Yvon, 'Analogical learning and formal proportions: Definitions and methodological issues', Technical Report ENST-2005-D004, École Nationale Supérieure des Télécommunications, (June 2005).