*TALN 2010, Montréal, 19–23 juillet 2010*

# Tree analogical learning. Application in NLP

A.Ben Hassena L.Miclet
ENSSAT / IRISA, Lannion, France
{benhasse, miclet}@enssat.fr

**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 presents some of its uses in machine learning for NLP. The analogical proportion is a relation between four objects that 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.
We firstly give definitions about the general notion of analogical proportion between four objects. We give a special focus on objects structured as ordered and labeled trees, with an original definition of analogy based on optimal alignment. Secondly, we present two algorithms which deal with tree analogical matching and solving analogical equations between trees. We show their use in two applications : the learning of the syntactic tree (parsing) of a sentence and the generation of prosody for synthetic speech.

**Résumé.** En intelligence artificielle, l'analogie est utilisée comme une technique de raisonnement non exact pour la résolution de problèmes, la compréhension du langage naturel, l'apprentissage des règles de classification, etc. Cet article s'intéresse à la proportion analogique, une forme simple du raisonnement par analogie, et présente son application en apprentissage automatique pour le TALN. La proportion analogique est une relation entre quatre objets qui exprime que la manière de transformer le premier objet en le second est la même que la façon de transformer le troisième en le quatrième.
Premièrement, nous définissons formellement la proportion analogique entre quatre objets. Nous nous intéressons particulièrement aux objets structurés que sont les arbres ordonnés et étiquetés, avec une définition originale de l'analogie fondée sur l'alignement optimal. Ensuite, nous présentons deux algorithmes qui calculent la dissemblance analogique entre quatre arbres et qui trouvent des solutions, éventuellement approchées, à une équation analogique entre arbres. Nous montrons leur utilisation dans deux applications : l'apprentissage de l'arbre syntaxique d'une phrase et la génération de la prosodie dans la synthèse de parole.

**Mots-clés :** Proportion analogique, arbre syntaxique, analyseur syntaxique analogique.

**Keywords:** Analogical proportion, syntactic tree, analogical syntactic parser.

## 1 Introduction

The concept of analogy has been studied as one of the modality of reasoning since Aristotle ((Lepage, 2003) ; (Holyoak, 2005)). It is a form of reasoning by generalization, but neither abductive nor inductive, which models a third form of learning. Recently, a growing interest has been manifested for a formal

point of view on the analogical proportion. This concept is now rigorously defined and its applications in representation spaces of various kinds have been developed, with interesting operational results. Its application to learning and generation is conceptually simple ; however, as in many areas of artificial intelligence, the complexity of some algorithmic problems remains 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 extend the concept of alignment defined by (Jiang *et al.*, 1994) to 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. In the third section, we describe our approach, starting from several original definitions to define the analogical dissimilarity between trees. Section 4 presents 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 of a sentence and give hints to use it to predict prosody in speech synthesis.

# 2 The Analogical Proportion and its Applications

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 as : "$O_1$ is to $O_2$ as $O_3$ is to $O_4$" and is denoted by $O_1 : O_2 :: O_3 : O_4$ (Lepage, 2003).

## 2.1 The Analogical Proportion, an algebraic definition

**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. $A : B :: C : D \Leftrightarrow A : C :: B : D$

*Moreover, every couple of elements must satisfy the relation : $A : B :: A : B$*

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.

A nice formulation of the analogical proportion according to these axioms, based on the notion of factorisation, has been given in (Stroppa & Yvon, 2005).

## 2.2 Semantics and examples

The relation $A : B :: C : D$ is generally interpreted as "$A$ is to $B$ as $C$ is to $D$", which means that to transform $A$ into $B$, one has to make the same operations than to transform $C$ in $D$. The analogical proportion has been studied and its semantics explored especially in the following domains :

$\{0,1\}^d$**.** Four objects defined by a vector of binary attribute are in analogical proportion when, on each coordinate, one has one among the six proportions : $0:0::1:1, 1:1::0:0, 1:0::1:0, 0:1::0:1$, $0:0::0:0$ or $1:1::1:1$.

$\mathbb{R}^d$**.** Four objects defined as vectors of numerical attributes are in additive analogical proportion when $\overrightarrow{AC} = \overrightarrow{BD}$, i.e. on every coordinate : $A_i : B_i :: C_i : D_i \Leftrightarrow A_i + D_i = B_i + C_i$ .

$\Sigma^\star$**.** Some theory and algorithms about the analogical proportion on strings and its use can be found in (Lepage, 2003; Stroppa & Yvon, 2005; Miclet *et al.*, 2008; Hofstadter, 1994; Langlais *et al.*, 2008). An analogical proportion on strings is, for example : `overlook` is to `looked` as `overcook` is to `cooked`.

# 3 Analogical proportion and trees

With the same principles that (Stroppa & Yvon, 2005) and (Miclet *et al.*, 2008) for strings, we present in this paragraph our methodology to define an analogical proportion between ordered and labelled trees. The only prerequisite is that there exists an analogical proportion in the alphabet of the label of the nodes, augmented with the empty word $\lambda$. A node with label $\lambda$ is also denoted $\lambda$ and is called an *empty node*.

**Definition 3.1** *(**Alignment between two trees (Jiang** **et al., 1994**)) An alignment between two trees $T_1$, $T_2$ whose labels are in $\Sigma_\lambda = \Sigma \cup \lambda$ is a tree with labels in $(\Sigma_\lambda) \times (\Sigma_\lambda)/(\lambda, \lambda)$ which first projection is $T_1$, where the empty nodes $\lambda$ are ignored and which second projection is $T_2$, where the empty node $\lambda$ 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.

**Definition 3.2** *(**Analogical proportion between trees**) Let $x$, $y$, $z$ and $t$ be four trees whose labels are in $\Sigma_\lambda$. We suppose that an analogical proportion exists in $\Sigma_\lambda$. 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 $\Sigma_\lambda^4$, 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 (figure 1 (a),(b)).*

**Definition 3.3** *(**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.*

## 3.1 Analogical Dissimilarity between trees

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*.

This measure has been introduced on binary vectors, numerical vectors and sequences in (Miclet *et al.*, 2008), and we want now to extend it to trees. We present in the following our definition of AD between four trees and we give its properties. This definition uses the notion of alignment between four trees. We assume there is some analogical dissimilarity existing on the alphabet of the nodes labels $\Sigma_\lambda$. Let $a$, $b$, $c$ and $d$ be node labels in $\Sigma_\lambda$. We have $AD(a,b,c,d) = 0$ iff $a : b :: c : d$ and $AD(a,b,c,d) > 0$ in the other case. Thus, the Analogical Dissimilarity between four ordered labeled trees can be defined by :

**Definition 3.4** ***(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 $A$, and the cost (or analogical dissimilarity) of an alignment is defined as :*

$AD(X,Y,Z,T) = \sum AD(x_i, y_i, z_i, t_i)$*, with* $i \in [1..|A|]$ *and* $x_i$*,* $y_i$*,* $z_i$ *and* $t_i \in \Sigma_\lambda$*.*

**Definition 3.5** ***(Best 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 :*

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

## 3.2 Algorithms

We have described in (BenHassena & Miclet, 2010) two algorithms based on dynamic programming, called *AnaTree* and *SolvTree* that allow :

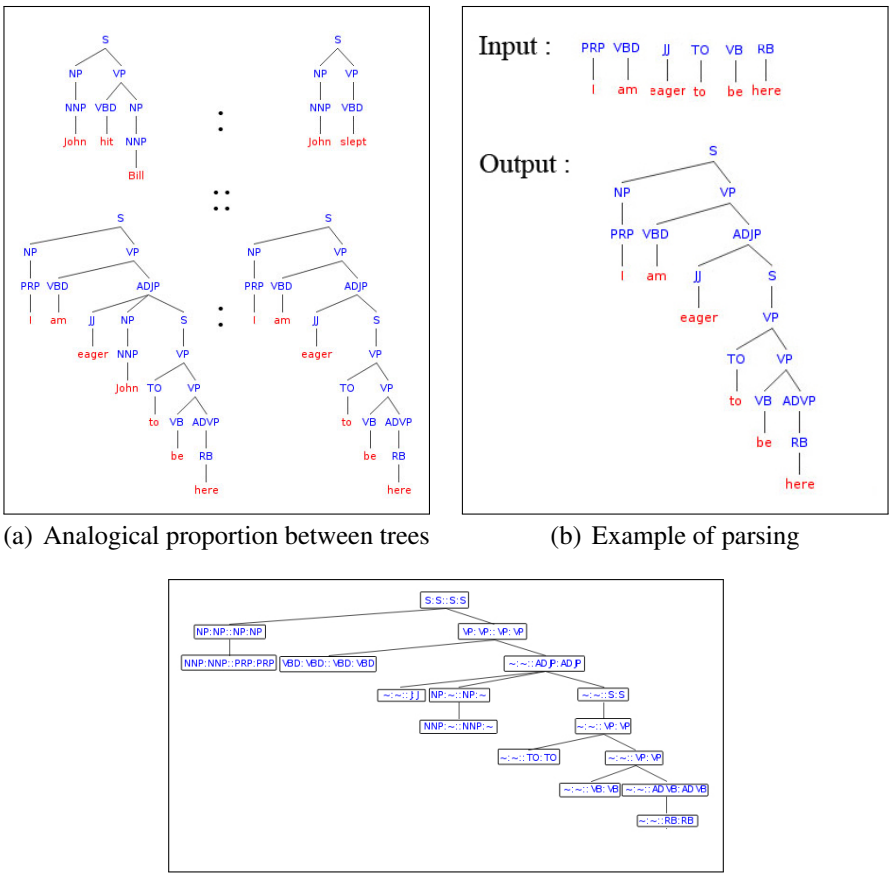
– from four trees, to compute their $AD$ and their optimal alignment, according to definition 3.4 ; this procedure is in $O(|T|^4)$, where |T| is the number of nodes of the bigger tree.
– from three trees $T_1$, $T_2$ and $T_3$, to compute a fourth tree $T_4$ such that $AD(T_1, T_2, T_3, T_4)$ is minimal, according to definition 3.5. This procedure is in $O(|T|^3)$.

# 4 Analogical syntactic parser

We apply these algorithms of analogical matching between structured objects to automatic parsing. Our hypothesis is simple : if four sentences have their sequences in analogical proportion (or with a low $AD$), then their syntactic trees are also in analogical proportion (or with a low $AD$). We test this hypothesis on a corpus of parsed sentences, each of these sentences being associated to its syntactic tree. The sentences are composed as sequences of grammatical categories, which are found as labels of the syntactic tree leaves. The labels of the tree nodes are syntactic labels as the "noun phrase (NP)", etc. Figure 1 (b) shows an example of parsing.

This assumption leads to a method for automatically generating syntactic tree (parsing automatic). 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 analogical proportion $S_0 : S_1 : S_2 : S_3$




(a) Analogical proportion between trees

(b) Example of parsing

(c) The tree resulting of an analogical proportion

FIGURE 1 – Analogical syntactic parser

2. We make the hypothesis 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 (Marcus *et al.*, 1993). Since the data available is limited, we have used the leave-one-out technique to evaluate the results. Preliminaries results give an exact or almost exact (with a AD lower than 2) restitution of the parsing tree from the sequence in 82 % of cases. Note that, since we have kept in the learning set only sentences that have different sequences of grammatical categories, a simple nearest neighbor method would have given an null accuracy at a null distance. This shows that the analogical proportion takes profit of the recombination of subsequences and of subtrees.

# 5 Analogical prosody prediction

We present here ongoing work on prosody prediction for speech synthesis. This approach considers sentences as tree structures and infers the prosody from a corpus of such structures using a machine learning technique. The prediction is achieved from the prosody of the closest (at minimal AD) sentences of the corpus through tree analogical dissimilarity measurements, using also the analogy-based approach. Given three known tree structures $T_1$, $T_2$, $T_3$ and a new one $T_0$, an analogical proportion would be expressed

as : $T_1$ is to $T_2$ as $T_3$ is to $T_0$ if and only if the set of operations transforming $T_1$ into $T_2$ is equivalent to the set of operations transforming $T_3$ into $T_0$. This relation can be relaxed according to the notion of analogical dissimilarity. Next, the analogical transfer would apply on the prosodic string associated to $T_3$ the transformation defined between the prosodic strings associated to $T_1$ and $T_2$ to produce the prosody of $T_0$. From these two notions, the analogical inference would be therefore defined as :

– firstly, retrieve all analogical proportions involving $T_0$ and three trees in the corpus ;
– secondly, compute the analogical transfer for each 3-tuple of the corresponding prosodic strings, and store the result in a set of possible outputs if the transfer succeeds.

Experiments are currently under process to qualify this approach.

# 6 Conclusion

In this paper, we have proposed a new learning method based on the analogical proportion between trees, and its approximation called "analogical dissimilarity". As an application we have firstly proposed a case-based parsing system, with preliminary promising results. Secondly, we have presented hints on a new prosody prediction method. Further investigations will be conducted, taking into account that the proposed brute force method is time-consuming, but can be ameliorated, as shown in (Langlais & Yvon, 2008) and (Miclet *et al.*, 2008).

# References

BENHASSENA A. & MICLET L. (2010). Analogical learning using dissimilarity between tree-structures. In *Proceedings of ECAI 2010*. To be published.

HOFSTADTER D. (1994). *Fluid Concepts and Creative Analogies*. New York : Basic Books.

HOLYOAK K. (2005). Analogy. In *The Cambridge Handbook of Thinking and Reasoning*, chapter 6. Cambridge University Press.

JIANG T., WANG L. & ZHANG K. (1994). Alignment of trees - an alternative to tree edit. In *CPM '94 : Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching*, p. 75–86, London, UK : Springer-Verlag.

LANGLAIS P. & YVON F. (2008). Scaling up analogical learning. In *22nd International Conference on Computational Linguistics (COLING 2008)*, p. 51–54, Manchester, United Kingdom. Poster.

LANGLAIS P., YVON F. & ZWEIGENBAUM P. (2008). *An Analogical Learning Approach to Translating Terms*. Rapport interne, Paritech, INFRES, IC2, Paris, France.

LEPAGE Y. (2003). *De l'analogie rendant compte de la commutation en linguistique*. Grenoble. Habilitation à diriger les recherches.

MARCUS M. P., MARCINKIEWICZ M. A. & SANTORINI B. (1993). Building a large annotated corpus of english : the penn treebank. *Comput. Linguist.*, **19**(2), 313–330.

MICLET L., BAYOUDH S. & DELHAY A. (2008). Analogical dissimilarity : Definition, algorithms and two experiments in machine learning. *Journal of Artificial Intelligence Research*, **32**, 793–824.

STROPPA N. & YVON F. (2005). *Analogical learning and formal proportions : Definitions and methodological issues*. Rapport interne ENST-2005-D004, École Nationale Supérieure des Télécommunications.