Analogical proportions in a lattice of sets of alignments built on the common subwords in a finite language
Laurent Miclet, Nelly Barbot, Baptiste Jeudy

To cite this version:
Laurent Miclet, Nelly Barbot, Baptiste Jeudy. Analogical proportions in a lattice of sets of alignments built on the common subwords in a finite language. 1st International Workshop on Similarity and Analogy-based methods in AI (SAMAI), Aug 2012, Montpellier, France. pp.25-31. hal-00760722

HAL Id: hal-00760722
https://hal.archives-ouvertes.fr/hal-00760722
Submitted on 4 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract. We define the locally maximal subwords and locally minimal superwords common to a finite set of words. We also define the corresponding sets of alignments. We give a partial order relation between such sets of alignments, as well as two operations between them. We show that the constructed family of sets of alignments has the lattice structure. The study of analogical proportion in lattices gives hints to use this structure as a machine learning basis, aiming at inducing a generalization of the set of words.

Keywords: Locally maximal subwords, alignments, algebraic structure of sets of alignments on a set of words (lattice), analogical proportion.

1 Introduction

Much has been done on finding maximal subwords and minimal superwords to a set of words, when the order relation is based on the length of words. We are interested in this paper in the same problem, but for the finer order relation based on the definition of a subword. Is there a manner to characterize the set of maximal subwords and that of minimal superwords, given a finite set of words, according to this relation? More than that, is there an algebraic relation between all these sets of subwords and superwords of a word? An answer to these questions would allow to give a precise definition to what the words of a set share, and how this common core is organised.

The first part of this paper gives a partial answer to these points. We define in section 2 a particular case of the notion of alignment, which will be useful for our construction. Actually, in section 3, we define two operations and an order relation on sets of alignments that leads to the construction of a lattice.

We are also interested in how this structure could be analysed in terms of analogical proportions, which could be used in machine learning. Since we start from a finite set of words, the convenient machine learning framework seems to be grammatical inference (from a positive set of positive samples, in our case). It seems that the lattice structure is partially adapted to learning by analogy, since some natural analogical proportions can be observed in such a structure. We give in section 4 some hints on these points.

2 Maximal subword, minimal superword, alignment

2.1 Basics

Let be an alphabet, i.e. a finite set of letters. A word is a sequence of letters in . The length of , denoted , is . The empty word, of null length, is . A language is a set of words. A subword of a word is a word obtained by deleting letters of at some (non necessarily adjacent) positions in . We denote , the shuffle of the two words and .

In , the set of all words on , we use the order relation defined by: ( is a subword of ). When is a subword of , is called a superword of . For example: .

A word is a common subword to and if there does not exist any other common subword to and such that . For example, and are maximal common subwords to such that .

Defining a common maximal subword to a finite set of words is a straightforward extension.

A maximal common subword to two words and to a non empty finite set of words is defined in an analogous way.

In a partially ordered set , an antichain is a subset of composed of pairwise incomparable elements. Any subset of can be reduced to a maximal antichain by removing from every element of less than another element of .

2.2 Alignments

2.2.1 Definition

Definition 1 An alignment is a finite set of pairs where is a word and is a set of indices between and . The set defines a subword of denoted . Moreover, an alignment must satisfy the following properties for all , , , and :

1. 
2. 
3. 

The set of words on which the alignment is defined is called the support and is denoted .

5 Other terms for subword are subsequence and partial word. A factor, or substring is a subword of built by contiguous letters of .

A superword of , also called a supersequence must not be confused with a superstring of , in which the letters of are contiguous. In other words, is a factor (a substring) of any superstring of . See [Gus97], pages 309 and 426.
An alignment (regardless of the third point of our definition), is called a trace by Wagner and Fisher [WF74] for two words and a threading scheme in Maier [Mat78].

According to our definition\(^6\), the support is an antichain of words for \(\leq\).

The set of indices \(l\) will be called the position of the indexed subword of \(w[l]\).

In the following, an alignment will be represented by a set of words in which some letters are boxed. For each element \((w, l)\) of the alignment, the boxed letters represent the subword \(w[l]\) (also called the boxed subword of the alignment).

For legibility, the \(n\) words can be displayed in such a manner that the corresponding letters of \(w\) in the \(n\) words are in the same column. Some blanks can be added freely to help the reading. For example:

\[
\begin{align*}
a &= \left( \begin{array}{cccccc}
\alpha & \alpha & \beta & \beta & \beta & \beta \\
\beta & \beta & \gamma & \gamma & \gamma & \gamma \\
\delta & \delta & \delta & \delta & \delta & \delta \\
\end{array} \right)
\end{align*}
\]

denotes the alignment

\[
a = \{(acbdeg, \{1, 2\}), (aceh, \{1, 2\}), (ghedc, \{2, 4\})\}
\]

We can write also without ambiguity:

\[
a = \{(a b c)\}_{\text{deg}}, (a c h), (g a h e c)\}
\]

### 2.2.2 Locally maximal alignments and locally maximal subwords

Generally speaking, two alignments on the same support \(W = \{w_1, \ldots, w_n\}\) with the same boxed subword \(v\) can be different (having different set of indexes). We could define maximal alignments as those whose boxed letters are maximal subword of \(W\).

However, all interesting alignments would not be maximal with this definition. For example the two words \(w_3 = abed\) and \(w_2 = dababc\). The complete set of common subwords is \(\{e, a, b, c, ab, ac, bc, abc, d\}\) and their set of maximal common subwords is \(\{abc, d\}\).

But these two subwords are not sufficient to define the totality of the interesting alignments. Actually the alignment \((a b c)\}_{\text{deg}}, (a c h), (g a h e c)\) is somehow "maximal" since it is not comparable to the only alignment with the boxed subword \(abc\), namely \((a b c)\}_{\text{deg}}, (a c h), (a b c)_{abc}\).

This leads to define the following notion of locally maximal alignment and of locally maximal subword.

**Definition 2** An alignment \(a = \{(w_1, l_1), \ldots, (w_n, l_n)\}\) is locally maximal if there is no other alignment \(b = \{(w_1', l_1'), \ldots, (w_n', l_n')\}\) on the same support such that for all \(i, l_i \subset l_i'\).

Notice that the empty alignment \(\emptyset\) is locally maximal.

**Definition 3** The set of boxed subwords associated to all locally maximal alignments between a finite set of words \(W = \{w_1, \ldots, w_n\}\) is called the set of locally maximal subwords to \(W\) and is denoted \(\cap(W)\).

For some \(v \in \cap(W)\), the set of locally maximal alignments associated to \(v\) is denoted \(A_v(W)\).

We also define: \(A(W) = \bigcup_{v \in \cap(W)} A_v(W)\).

For example, let us consider \(W = \{ababc, cabcd\}\), its sets of locally maximal alignments are given by

\[
A_{ab}(W) = \{\{a b c \}_{\text{deg}}, \{a c a \}_{\text{bde}}, \{a c c \}_{\text{bde}}\},
\]

\[
A_{ac}(W) = \{\{a b b \}_{\text{cde}}, \{a b d \}_{\text{cde}}\},
\]

\[
A_{bc}(W) = \{\{a b c \}_{\text{d}}, \{a c b \}_{\text{d}}\}
\]

Let us consider \(W = \{\{a b c \}_{\text{d}}, \{a c b \}_{\text{d}}\}\). By definition of \(L(A_{W}(\{u, v\}))\), there exists \((u^1, \ldots, u^{p+1})\) and \((v^1, \ldots, v^{p+1})\), respectively sets of factors of \(u\) and \(v\), such that the word \(W\) can be written as \(W = u^1w_1 \ldots u^pw_{p+1}v^{p+1}v^1\), where, for every \(i \in \{1, \ldots, p + 1\}\), \(x^i \in \{u^1 \cdot v^1\}\).

### 2.3 Language associated with an alignment

**Definition 4** Let \(w = w_1 \cdots w_p\) be a word, locally maximal subword of two words \(u\) and \(v\) at only one position (i.e. \(|A_{W}(\{u, v\})| = 1\)). Then there exists an unique set of factors of \(u\), denoted \((u^1, \ldots, u^{p+1})\), and an unique set of factors of \(v\), denoted \((v^1, \ldots, v^{p+1})\), such that \(u = u^1w_1 \ldots u^pw_{p+1}\) and \(v = v^1w_1 \ldots v^{p+1}w_{p+1}v^{p+1}\). We define \(L(A_{W}(\{u, v\}))\) as the following finite language:

\[
L(A_{W}(\{u, v\})) = \{u^1v^1w_1(u^2v^2)^2, \ldots, (u^p \cdot v^p)w_p(u^{p+1} \cdot v^{p+1})\}
\]

The construction of \(L(A_{W}(\{u, v\}))\) is shown in Figure 1, with straightforward graphic conventions.

![Figure 1. The construction of \(L(A_{W}(\{u, v\}))\) when \(|A_{W}(\{u, v\})| = 1\).](image)

If \(|A_{W}(\{u, v\})| > 1\), \(L(A_{W}(\{u, v\}))\) is defined as the union of all languages associated with all different positions of \(w\) as locally maximal subword of \(u\) and \(v\). Finally, \(L(A(u, v))\) is defined as the union of the languages \(L(A_{W}(\{u, v\}))\), for all \(w\) locally maximal subwords of \(u\) and \(v\).

**Proposition 1** Let \(w\) be a locally maximal subword common to two words \(u\) and \(v\) and \(L(A_{W}(\{u, v\}))\) constructed as above. We have:

1. All words in \(L(A_{W}(\{u, v\}))\) are (non necessarily minimal) common superwords of \(u\) and \(v\).
2. For any word \(W \in L(A_{W}(\{u, v\}))\), we have\(^7\) \(|W| + |w| = |u| + |v|\).

**Proof.** We firstly give an example to show that a word of \(L(A_{W}(\{u, v\}))\) can be a non-minimal superword of \(u\) and \(v\).

We take the two words \(w = abacbb\) and \(v = aabbc\). The associated alignment \((a b a \cdot (a b) \cdot c)\) is locally maximal. The language \(L(A_{a b a c b a c b c})(\{u, v\})\) contains the language \(a a a b a c b c\) and, in particular, the word \(w = abacbc\). The word \(w' = abacbc\) is another superword of \(u\) and \(v\), and \(w' \leq w\). Thus, \(w\) is not an locally maximal superword of \(u\) and \(v\).

Then we demonstrate the proposition.

Let us consider \(W \in L(A_{W}(\{u, v\}))\). By definition of \(L(A_{W}(\{u, v\}))\), there exists \((u^1, \ldots, u^{p+1})\) and \((v^1, \ldots, v^{p+1})\), respectively sets of factors of \(u\) and \(v\), such that the word \(W\) can be written as \(W = x^1w_1 \ldots x^pw_{p+1}v^{p+1}v^1\), where, for every \(i \in \{1, \ldots, p + 1\}\), \(x^i \in \{u^i \cdot v^i\}\).

\(^7\) A consequence of this assertion is: let \(LCS(u, v)\) be a longest common subword to \(u\) and \(v\) and \(SCS(u, v)\) be a shortest common superword to \(u\) and \(v\). Then we have: \(|LCS(u, v)| + |SCS(u, v)| = |u| + |v|\).
1. Therefore, for every \( i \in \{1, \ldots, p + 1\} \), \( x^i \geq u^i \) and \( x^i \geq v^i \).

We then have \( W \geq u^1 w_1 \ldots u^p w_p v^{p+1} = u \) and \( W \geq v^1 w_1 \ldots v^p w_p v^{p+1} = v \).

2.

\[
|W| = |x^1| + |w_1| + \ldots + |x^p| + |w_p| + |x^{p+1}|
= |w| + \sum_{i=1}^{p+1} |x^i| = |w| + \sum_{i=1}^{p+1} (|u^i| + |v^i|)
= |w| + (|u| - |w|) + (|v| - |w|) = |u| + |v| - |w|.
\]

2.4 Constructive algorithms

We have devised an algorithm producing a finite automaton \( A_{\cap\{u,v\}} \) which exactly recognizes the language \( \cap\{u,v\} \), the set of locally maximal subwords common to two words \( u \) and \( v \), due to lack of space, we do not describe it here. It is based on the transformation of an 2-d array displaying which letters are common to two words into a finite automaton recognizing \( \cap\{u,v\} \) (see an example on figure 2(a)).

Starting from \( A_{\cap\{u,v\}} \), it is then simple to produce a finite automaton that we call \( A_{\cup\{u,v\}} \) which exactly recognizes the language \( L(A_{\{u,v\}}) \) (also denoted \( \cup\{u,v\} \)). We display an example at figure 2(b).

3 Order relation and operations between alignments

In this section, we are interested in a particular family of alignments, since we want to describe what have in common the subwords and superwords of a finite set \( U \) of sentences. We will consider alignments on \( U \), i.e. alignments with a support subset of \( U \). Moreover, we will assume that \( U \) is an antichain according to the order relation \( \subseteq \).

3.1 Order relation

**Definition 5 (Order on alignments on \( U \))**

Given two alignments on \( U \) \( a = \{(w_1, l_1), \ldots, (w_n, l_n)\} \) and \( b = \{(w'_1, l'_1), \ldots, (w'_m, l'_m)\} \), we write \( a \subseteq b \) if for all \( i \in (1, n) \), it exists \( j \in (1, m) \) such that

1. \( w_i = w'_j \)
2. \( l'_j \subseteq l_i \)

Therefore, if \( a \subseteq b \), then \( \text{word}(a) \subseteq \text{word}(b) \).

It is easy to check that \( \subseteq \) is a partial order relation on the set of alignments and that the empty alignment \( \emptyset \) is smaller than every other alignment.

**Definition 6 (Homogeneous sets of alignments)**

A set of alignments is homogeneous if it is non empty and all its elements have the same support. The family of homogeneous sets of locally maximal alignments is denoted \( A_H \).

In order to link this definition with definition 3, we can notice that, for any subset \( W \) of \( U \), \( A(W) \in A_H \).

**Definition 7 (Order on homogeneous sets of alignments on \( U \))**

Let \( A \) and \( B \) be two homogeneous sets of alignments. We have \( A \subseteq B \) if for all \( b \in B \), there is \( a \in A \) such that \( a \subseteq b \).

**Proposition 2** \( \subseteq \) is a partial order on \( A_H \) and the smallest element is \( \{\} \).

**Proof.**

Reflexivity and transitivity are immediate. In order to check the antisymmetry, let us consider two homogeneous sets of locally maximal alignments, denoted \( A \) and \( B \), such that: \( A \subseteq B \) and \( B \subseteq A \). Since \( A \) and \( B \) are homogeneous, all alignments in \( A \) have the same support, denoted \( \text{word}(A) \), and the same holds for \( B \), with the support denoted \( \text{word}(B) \). From the definition of \( \subseteq \), we easily check that \( \text{word}(A) = \text{word}(B) \). Let us consider \( b_1 = \{(w_1, l_1), \ldots, (w_n, l_n)\} \in B \); since \( A \subseteq B \) and \( B \subseteq A \), it exists \( a \in A \) and \( b_2 \in B \) such that \( a \subseteq b_1 \) and \( b_2 \subseteq a \). By transitivity, we have \( b_2 \subseteq b_1 \). At last, \( b_1 \) and \( b_2 \) having the same support and being locally maximal, it implies that \( b_1 = b_2 \) and then \( a \in B \). Hence, \( A \subseteq B \). Similarly, we can check that \( B \subseteq A \).
3.2 Definition and properties of $\bigodot$

Definition 8 Let $a \in A_r(\{u_1, \ldots, u_n\})$ and $b \in A_r(\{v_1, \ldots, v_m\})$, where $a = \{(u_1, l_1), \ldots, (u_n, l_n)\}$ and $b = \{(v_1, l'_1), \ldots, (v_m, l''_m)\}$. Firstly, we construct $a \bigodot b$, the finite set of alignments $c = \{(w_1, l_1), \ldots, (w_p, l_p)\}$ such that

1. $\{w_1, \ldots, w_p\} = \text{word}(a) \cup \text{word}(b)$
2. for all $(i, k)$, if $(w_k = u_i)$ then $(L_k \subseteq l_i)$
3. for all $(j, k)$, if $(w_k = v_j)$ then $(L_k \subseteq l'_j)$

Secondly, we denote $a \bigodot b$ the set of minimal elements of $a \bigodot b$ according to $\sqsubseteq$.

As consequence, if $\cap \{(r, s)\} \neq \emptyset$, then the boxed word in $c \in a + b$ is a subword of $r$ and $s$, else, no letter is boxed in $c$. In addition, if $a$ and $b$ contains an identical word $u_i = v_j$ such that $l_i \cap l'_j = \emptyset$, no letter is then boxed in $c$.

The operation $\bigodot$ is extended to homogeneous sets of alignments by the following definition.

Definition 9 Let $A$ and $B$ be two homogeneous sets of alignments. We define $A \bigodot B$ as the set of the minimal elements of $A + B$ according to $\sqsubseteq$ where

$$A + B = \bigcup_{a \in A} a \bigodot B$$

Proposition 3 The operation $\bigodot$ is internal to $A_H$, commutative and idempotent.

Proof. Let us consider $A \in A_H$ and $B \in A_H$.

1. All the alignments in $A \bigodot B$ are locally maximal by definition and have the same support, namely $\text{word}(A) \cup \text{word}(B)$.
2. The commutativity is straightforward.
3. Let $a$ be an element of $A$, it is immediate that $a \in (a + a) \subseteq A + A$. Moreover, since $A \in A_H$, $a$ is a locally maximal alignment, and so $a \in A \bigodot A$. Consequently, $A \subseteq A \bigodot A$. Reciprocally, let $c$ be an element of $A \bigodot A$. Then it exists a couple $(a, b) \in A^2$ such that $c \in a + b$. Since $A \in A_H$ and $\text{word}(c) = \text{word}(a) \cup \text{word}(b)$, $a$ and $b$ have the same support. Moreover, from definitions 5 and 8, $a \subseteq A$ and $b \subseteq A$ being a minimal element of $A + A$ according to $\sqsubseteq$, and $a$ and $b$ belonging to $A + A$, it turns out that $a = b = c$. At last, $c \in A$. Hence $A \bigodot A \subseteq A$. If $\bigodot$ is then idempotent on $A_H$. □

3.3 Construction of $\bigodot$

Definition 10 Let $a \in A_r(\{u_1, \ldots, u_n\})$ and $b \in A_r(\{v_1, \ldots, v_m\})$ where $a = \{(u_1, l_1), \ldots, (u_n, l_n)\}$ and $b = \{(v_1, l'_1), \ldots, (v_m, l''_m)\}$. We construct a $\bigodot b$, the finite set of alignments $c = \{(w_1, l_1), \ldots, (w_p, l_p)\}$ such that

1. $\{w_1, \ldots, w_p\} = \text{word}(a) \cap \text{word}(b)$
2. Either, for all $(i, k)$ such that $w_k = u_i$ we have $l_i \subseteq L_k$, or for all $(j, k)$ such that $w_k = v_j$ we have $l'_j \subseteq L_k$.
3. $c$ is a locally maximal alignment.

An alignment in a $\bigodot b$ is thus based either on a restriction of $a$ to the support $\text{word}(a) \cap \text{word}(b)$ or on a restriction of $b$ to the same support. For instance, if $a = \{a_d, ab, a_c, a_b\}$ and $b = \{a_c, b, ab, c, c_b\}$, then $a \bigodot b = \{ab, a_b, b, c, c_b\}$.

Definition 11

$$A \bigodot B = \bigcup_{a \in A \cap B} (a \bigodot b)$$

Proposition 4 The operation $\bigodot$ is internal to $A_H$, commutative and idempotent.

Proof. The commutativity is straightforward (definition 10 is symmetric wrt $a$ and $b$). For idempotence, we use the fact (direct consequence of the definition) that if $a$ and $b$ are locally maximal alignments on the same support, then $a \bigodot b = \{a, b\}$. Let us consider $A \subseteq A_H$; if $a \in A$ then $a \bigodot (a \bigodot A) \subseteq (a \bigodot A)$ and therefore $A \subseteq A \bigodot A$. If $c \in A \bigodot A$, then there exists $(a, b) \in A^2$ such that $c \in a \bigodot b$. Since $a$ and $b$ have the same support, either $c = a$ or $c = b$, therefore $c \in A$ and $A \bigodot A \subseteq A$. □

3.4 Structure of homogeneous sets of alignments on $U$

We define $\text{sup}(A, B)$ as the minimal set of alignments larger than $A$ and $B$ (if it exists) according to $\sqsubseteq$. Similarly, $\text{inf}(A, B)$ is the maximal set of alignments smaller than $A$ and $B$.

Proposition 5 Let $A$ and $B$ be finite homogeneous sets of alignments. Then $\text{sup}(A, B)$ exists and:

$$\text{sup}(A, B) = A \bigcup B$$

Proof. First, we show that $A \bigcup B$ is greater than $A$ and $B$ for $\sqsubseteq$. Let $c \in A \bigcup B$. By construction, there exist $a \in A$ and $b \in B$ such that $c \in a \bigcup b \subseteq a + b$. By the first item of definition 8, $\text{word}(c) \subseteq \text{word}(a) \cup \text{word}(b)$ and by the two other items, we can conclude that $a \sqsubseteq c$. Thus for every $c \in C$ there is $a \in A$ such that $a \sqsubseteq c$. Hence $A \subseteq A \bigcup B$ and $B \subseteq A \bigcup B$.

Let $C$ be a set of alignments greater than $A$ and $B$, and let $c \in C$. There are $a \in A$ and $b \in B$ such that $c \subseteq a \bigcup b \subseteq a + b$. By the second item of definition 8, $\text{word}(c) \subseteq \text{word}(a) \cup \text{word}(b)$ and by the two other items, we can conclude that $c \sqsubseteq a \bigcup b$. Remove from the support of $c$ all words not in the support of a or b. The obtained alignment may not be locally maximal, so we add more boxed letters to make it locally maximal. The result alignment $c'$ satisfies all conditions of Definition 8, thus $A \bigcup B \sqsubseteq C$ and therefore $\text{sup}(A, B) = A \bigcup B$.

There is no equivalent relation between $\bigodot$ and $\text{inf}$ for all homogeneous sets of alignments, we must restrict to sets of all alignments built on a given set of words.

Definition 12 If $U$ is a finite collection of words, we define the collection of sets of alignments $A(U) = \{A(V) | V \subseteq U\}$.

Proposition 6 Let $A$ and $B$ be sets of alignments in $A(U)$. Then, in $A(U)$, $\text{inf}(A, B)$ exists and:

$$\text{inf}(A, B) = A \bigodot B$$

Proof. First, we show that if $A = A(V)$ and $B = A(W)$ with $V \subseteq U$ and $W \subseteq U$ then $A \bigodot B = A(W \cap V)$. Let $c \in A \bigodot B$. $c$ is a locally maximal alignment on its support $\text{word}(A) \cap \text{word}(B) = W \cap V$, thus $c \in A(W \cap V)$. Let $c \in A(W \cap V)$. Let $a$ be an alignment on $W$ such that $c \subseteq a$, then $c$ is obtained from $a \in A$ using the definition of $A \bigodot B$ and $c \in A \bigodot B$. □
Let $C \in \mathcal{A}(U)$ be a set of alignments smaller than $A$ and $B$. We show that $C$ is smaller than $A \not\sqsubseteq B$. Some alignments of $C$ are smaller than alignments of $A$ and others are smaller than alignments of $B$. Since $C$ is homogeneous, its support $\text{word}(C)$ must be included in $\text{word}(A) \cap \text{word}(B)$ and since $C = A(T)$ for some $T \subseteq U$, then $T \subseteq V \cap W$. Therefore $A(T) \subseteq A(V \cap W)$ which is exactly $C \subseteq A \not\sqsubseteq B$.

**Proposition 7** Let $U = \{u_1, u_2, \ldots, u_n\}$ be a finite set of words, the operations $\sqsubseteq$ and $\sqsubseteq$ are internal to $\mathcal{A}(U)$.

**Proof.** For $\sqsubseteq$ it is a consequence of the previous definition. For $\sqsubseteq$, it is not difficult to see it from the definition of $\sqsubseteq$.

**Proposition 8** Let $U = \{u_1, u_2, \ldots, u_n\}$ be a finite set of words, antichain for $\sqsubseteq$. Then $U = (\mathcal{A}(U), \sqsubseteq, \sqsubseteq)$ is a lattice. This lattice is said to be built on the finite language $U$.

**Proof.** This is a direct consequence of the three previous propositions.

\section*{4 Analogical proportions in the lattice $\mathcal{U}$}

\subsection*{4.1 The axioms of analogical proportion}

**Definition 13** (Analogical proportion) An analogical proportion on a set $\Sigma$ is a relation in $\Sigma^4$ such that, for all 4-tuples $A, B, C$ and $D$ in relation in this order (denoted $A : B :: C : D$):

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

For every 2-tuple, one has: $A : B :: A : B$

It is easy to show that five other proportions are equivalent:

- $B : A :: D : C$
- $D : B :: C : A$
- $D : C :: B : A$
- $B : D :: A : C$
- $C : A :: D : B$

These requirements are often called the axioms of analogical proportion (see [Lep03]).

\subsection*{4.2 Analogical proportions between words}

**A first definition using factorization.** According to Yvon and Stroppa [SY05] a general definition of analogical proportion, conform to the axioms, can be given in many different cases thanks to the notion of factorization. We show here how it applies in $\Sigma^*$, and we will come back later to its use in general lattices.

**Definition 14** (Analogical proportions between words.)

$(x, y, z, t) \in \Sigma^*$ are in analogical proportion, which is denoted $x = y : z : t$, if and only if there exists a positive integer $n$ and two sets of words $(\alpha_i)_{i \in [1, n]}$ and $(\beta_j)_{j \in [1, n]} \in \Sigma^*$ such that:

- $x = \alpha_1 \ldots \alpha_n$, $t = \beta_1 \ldots \beta_n$, $y = \alpha_1 \beta_2 \alpha_3 \ldots \alpha_n$, $z = \beta_1 \alpha_2 \beta_3 \ldots \beta_n$
- $x = \alpha_1 \ldots \alpha_n$, $t = \beta_1 \ldots \beta_n$, $y = \beta_1 \alpha_2 \beta_3 \ldots \alpha_n$, $z = \alpha_1 \beta_2 \alpha_3 \ldots \beta_n$

and $\forall i, \alpha_i \beta_i \neq \epsilon$.

**Example.** reception : rejection :: deceptive : defective is an analogical proportion between sequences, with $n = 3$ and the factors:

- $\beta_1$ = re
- $\alpha_2$ = cept
- $\alpha_3$ = ion
- $\beta_2$ = de
- $\beta_3$ = fect
- $\beta_3$ = ive

<table>
<thead>
<tr>
<th>$\beta_1$</th>
<th>$\alpha_1$</th>
<th>$\beta_2$</th>
<th>$\alpha_2$</th>
<th>$\beta_3$</th>
<th>$\alpha_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>re</td>
<td>cept</td>
<td>ion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de</td>
<td>fect</td>
<td>ion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de</td>
<td>fect</td>
<td>ive</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The authors have shown that this definition is conform to the axioms.

**Another definition using alignments.** This second definition, with the associated algorithms, is given in [MBD08]. The axioms of analogical proportion are verified as well.

**Definition 15** Let $u, v, w$ and $x$ four words in $\Sigma^*$. We assume that an analogical proportion is defined on $\Sigma$. We extend this relations to $\Sigma_n = \Sigma \cup \{\epsilon\}$, adding the proportions $a : e :: a : e$ for all $a \in \Sigma$. Then $u, v, w$ and $x$ are in analogical proportion in $\Sigma^*$ if there exists an alignment between the four words such that every column of the alignment is an analogical proportion in $\Sigma_n$.

**Example.** Let $\Sigma = \{a, b, c, A, B, C\}$ an alphabet with the analogical proportions $a : b :: A : B$, $a : c :: A : C$, $b : c :: B : C$. The following alignment shows that there is an analogical proportion in $\Sigma^*$ between the four words $CaCA, CebBA, abAc$ and $bCbB$.

\[
\begin{pmatrix}
C & a & C & A \\
C & c & b & B & A \\
b & A & c \\
b & C & b & b
\end{pmatrix}
\]

Note that there is no boxed letter in this alignment. It can happen anyway in the case of a column such that $a : a :: a : a$.

**Links between the two definitions.** The second definition using alignments is shown to imply the first one (not the reverse). However, a straightforward modification of the first one lead to a complete equivalence [Has11].

\subsection*{4.3 Analogical proportions in a lattice}

Using the factorization technique, Stroppa and Yvon [SY05] have found that a general definition of an analogical proportion can be given in a lattice. Unfortunately, his definition was incomplete. We give here the complete one.

**Definition 16** For four elements $(x, y, z, t) \in (L, \vee, \wedge, \lambda)^4$, the analogical proportion denoted $(x : y :: z : t)$ is true if and only if:

- $x = (x \land y) \lor (x \land z)$ and $x = (x \lor y) \land (x \lor z)$
- $y = (x \land y) \lor (t \land y)$ and $y = (x \lor y) \lor (t \lor y)$
- $z = (t \land z) \lor (x \land z)$ and $t = (t \lor z) \lor (t \lor y)$
- $z = (t \land z) \lor (x \land z)$ and $z = (t \lor z) \lor (x \lor z)$

The geometry of this definition is displayed in figure 3.

A simple example of proportion in a lattice is given by the following property:
Let \( y \) be a subset of \( U \). Then the following analogical proportion holds:

\[
\forall m \in n, \exists n \in N \text{ such that } m \leq n.
\]

For example, \( \{ab,c\} \leq \{abcd,e\} \). There is an partial inclusion relation between the languages recognized by this lattice of automata, compatible with that of the subsets, since the following property holds.

**Proposition 11** For any subsets \( J \) and \( K \) of \( U \), the three following relations are equivalent: \( L(A(J)) \leq L(A(K)) \), \( J \subset K \) and \( \cap(K) \subset \cap(J) \).

Note that the exploration of such a lattice of automata, constructed on a finite set of positive examples, is the basis of the efficient finite automata inference, see \([\text{dH}10]\). This could be one basis for the use of our lattice in machine learning.

Another thread to follow could be the idea of analogical closure of a finite language, as described in \([\text{Lep}03]\) and that of analogical generation, see \([\text{BMMA}07]\). In both, a triple of words is taken in the learning sample and a fourth sentence is generated, under the constraint that the four sentences are in analogical proportion. It is not yet clear to the authors how this technique can be combined with the lattice structure, but this could be a connection with the area of machine learning on the basis of formal concepts, as in \([\text{Kuz}01]\).

**5 Conclusion and related work**

The problem of finding one longest common subsequence (subword) or one shortest common supersequence (superword) to two or more words has been well covered (see e.g. \([\text{Gus}97]\), pp 287-293 and 309, \([\text{IF92}]\)). However, to the best of our knowledge, the problem of finding an intentional definition to the sets of maximal subwords and minimal superwords of a set of words has not been explored yet. In this, we have produced, via the construction of a lattice of alignment sets, an interesting subset of minimal superwords and maximal subwords to a set of words. We have not worked yet neither on the theoretical complexity of the construction of the lattice of alignments, nor...
nor on its practical complexity and applications. Hereafter we give some bibliographical hints to this problem.

A complexity result (sometimes misinterpreted) is given by Maier [Mai78] who has demonstrated that the "yes/no longest common subsequence problem" and the "yes/no shortest common supersequence problem" are NP-complete for alphabets of sufficient size. These problems are defined as follows: "Given an integer \( k \) and a set of sequences \( R \), is \( |\text{LCS}(R)| \leq k \) ? Is \( |\text{SCS}(R)| \geq k \) ?" where \( |\text{LCS}(R)| \) and \( |\text{SCS}(R)| \) are the length of a longest common subsequence and the length of a shortest common supersequence of \( R \).

It is also true that finding the length of a shortest (longest) super(sub)sequence common to a set of \( k \) sequences is in \( O(m_1 \ldots m_k) \), with \( m_i \) the size of the \( i \)-th of the \( k \) sequences, hence exponential in \( k \).

The works of Fraser and Irving [FIM96] have produced algorithms to find the longest minimal common supersequence (superword) and the shortest maximal common subsequence, according to the order relation \( \leq \).

Yvon and Stroppa [SY05] give a definition of an analogical proportion between words and also within lattices. Our objective is to use the properties of the lattice structure on alignment sets to solve the associated analogical equations.

REFERENCES


9 The elementary operation is the comparison.