Journal of Machine Learning Research 21:164–176, 2012 The 11th ICGI

A Lattice of Sets of Alignments Built on the Common Subwords in a Finite Language

Laurent Miclet MICLET@ENSSAT.FR
IRISA-Dyliss (CNRS, INSA) & INRIA-Rennes, France.

Nelly Barbot BARBOT@ENSSAT.FR
IRISA-Cordial (CNRS, INSA), Lannion, France.

Baptiste Jeudy BAPTISTE.JEUDY@UNIV-ST-ETIENNE.FR
Laboratoire Hubert Curien (CNRS),
Université de Lyon à Saint-Étienne, France.

Editors: Jeffrey Heinz, Colin de la Higuera and Tim Oates

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. We give hints to use this structure as a machine learning basis for inducing a generalization of the set of words.

Keywords: Finite languages, locally maximal subwords, alignments, algebraic structure of sets of alignments on a set of words, lattice, machine learning, learning by analogy.

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 U of words, according to this relation? More than that, is there an algebraic relation between all these sets of subwords and superwords of U? An answer to these questions would allow to give a precise definition to what the words of U share, and how this common core is organised.

The first parts 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 useful in machine learning. Since we start from a finite set of words, the convenient machine learning framework seems to be grammatical inference (from a finite set of positive samples, in our case). It seems that the lattice structure is particularly 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 Σ be an alphabet, *i.e.* a finite set of letters. A *word* u is a sequence $u_1 \dots u_n$ of letters in Σ. The length of u, denoted $|u|$ is n. The empty word, of null length, is ϵ. A *language* is a set of words. A *subword* of a word u is a word obtained by deleting letters of u at some (non necessarily adjacent) positions[1] in u. We denote $u \bullet v$ the *shuffle* of the two words u and v.

In $\Sigma^\star$, the set of all words on Σ, we use the order relation $\leq$ defined by: ($u \leq v \Leftrightarrow$ u is a subword of v). When u is a subword of v, v is called a *superword*[2] of u. For example: $abc \leq aabbcd$.

A word w is a *common subword* to u and v when $w \leq u$ and $w \leq v$. The word w is a *maximal* common subword to u and v if there does not exist any other common subword x to u and v such that $w \leq x$. For example, ab and c are maximal common subwords to $u = cadba$ and $v = fagbhc$, while a is a non maximal common subword. Defining a common maximal subword to a finite set of words is a straightforward extension.

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

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

2.2. Alignments

2.2.1. Definition of Alignments

Definition 1 *An alignment is a finite set of pairs (w, l) where w is a word and l a set of indices between 1 and $|w|$. The set l defines a subword of w denoted $w[l]$. Moreover, an alignment* $\mathtt{a}$ *must satisfy the following properties for all $(w, l) \in \mathtt{a}$ and $(w', l') \in \mathtt{a}$:*

1. $w[l] = w'[l']$
2. $(w = w') \Rightarrow (l = l')$
3. $(w \leq w') \Rightarrow (w = w')$

The set of words on which the alignment is defined is called the *support* and is denoted $word(\mathtt{a}) = \{w \mid \exists l \subset \mathbb{N} \text{ with } (w, l) \in \mathtt{a}\}$. According to our definition[3], the support is an antichain of words for $\leq$.

1. Other terms for *subword* are *subsequence* and *partial word*. A *factor*, or *substring* is a subword of u built by contiguous letters of u.
2. A *superword* of u, also called a *supersequence* must not be confused with a *superstring* of u, in which the letters of u are contiguous. In other words, u is a factor (a substring) of any superstring of u. See (Gusfield, 1997), pages 4, 309 and 426.
3. An alignment (regardless of the third point of our definition), is called a *trace* by Wagner and Fisher (Wagner and Fisher, 1974) for two words and a *threading scheme* in Maier (Maier, 1978).

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:

$$\mathtt{a} = \begin{pmatrix} & \boxed{a} & & \boxed{c} & b & d & e & g & \\ & \boxed{a} & & \boxed{c} & & & e & & h \\ g & \boxed{a} & h & \boxed{c} & & d & & & \end{pmatrix}$$

denotes the alignment $\mathtt{a} = \{(acbdeg, \{1, 2\}), (aceh, \{1, 2\}), (gahcd, \{2, 4\})\}$. We can write also without ambiguity:

$$\mathtt{a} = (\boxed{a}\boxed{c}bdeg, \boxed{a}\boxed{c}eh, g\boxed{a}h\boxed{c}d).$$

2.2.2. Locally Maximal Alignments and Locally Maximal Subwords

Generally, two alignments on the same support $W = \{w_1, \ldots, w_n\}$ with the same boxed subword r 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. Consider for example the two words $w_1 = abcd$ and $w_2 = dabcab$. The complete set of common subwords is $\{\epsilon, a, b, c, ab, ac, bc, abc, d\}$ and their set of maximal common subwords is $\{abc, d\}$. Actually the alignment $(\boxed{a}\boxed{b}cd, dabc\boxed{a}\boxed{b})$ is somehow "maximal" since it is not comparable to the only alignment with the boxed subword abc, namely $(\boxed{a}\boxed{b}\boxed{c}d, d\boxed{a}\boxed{b}\boxed{c}ab)$.

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

Definition 2 *An alignment* $\mathtt{a} = \{(w_1, l_1), \ldots, (w_n, l_n)\}$ *is* locally maximal *if there is no other alignment* $\mathtt{b} = \{(w_1, l'_1), \ldots, (w_n, l'_n)\}$ *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* $\boxdot(W)$.

For some $r \in \boxdot(W)$*, the set of locally maximal alignments associated to* r *is denoted* $A_r(W)$.

We also denote $A(W) = \bigcup_{r \in \boxdot(W)} A_r(W)$.

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

$$\begin{aligned} A_{ab}(W) &= \{(\boxed{a}\boxed{b}abc, c\boxed{a}\boxed{b}d), (\boxed{a}ba\boxed{b}c, c\boxed{a}\boxed{b}d), (ab\boxed{a}\boxed{b}c, c\boxed{a}\boxed{b}d)\} \\ A_c(W) &= \{(abab\boxed{c}, \boxed{c}abd)\}. \end{aligned}$$

and $A(W) = A_{ab}(W) \bigcup A_c(W)$. Then, the set of locally maximal subwords of W is $\boxdot(W) = \{ab, c\}$.

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 $u = u^1 w_1 \ldots u^p w_p u^{p+1}$ and $v = v^1 w_1 \ldots v^p w_p v^{p+1}$. We define $L(A_w(\{u, v\}))$ as the following finite language:*

$$L(A_w(\{u, v\})) = (u^1 \bullet v^1)w_1(u^2 \bullet v^2), \ldots, (u^p \bullet v^p)w_p(u^{p+1} \bullet v^{p+1})$$

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

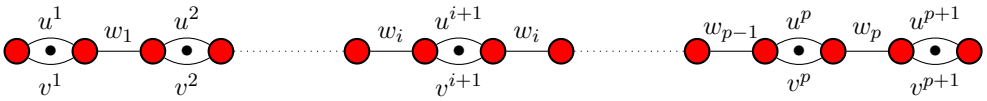


Figure 1: The construction of $L(A_w(u, v))$ when $|A_w(\{u\}, \{v\})| = 1$.

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 [4]) common superwords of u and v.*
2. *For any word $W \in L(A_w(\{u, v\}))$, we have[5] $|W| + |w| = |u| + |v|$.*

Proof.

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 $W = x^1 w_1 \ldots x^p w_p x^{p+1}$ where, for every $i \in \{1, \ldots p+1\}$, $x^i \in (u^i \bullet v^i)$.

1. Therefore, for every $i \in \{1, \ldots p+1\}$, $x^i \geq u^i$ and $x^i \geq v^i$.
   We then have $W \geq u^1 w_1 \ldots u^p w_p u^{p+1} = u$ and $W \geq v^1 w_1 \ldots v^p w_p v^{p+1} = v$.

4. Let two words $u = abcabb$ and $v = aabbc$, the associated alignment ($\boxed{\text{a}}\,\boxed{\text{b}}ca\boxed{\text{b}}b, \boxed{\text{a}}a\boxed{\text{b}}\,\boxed{\text{b}}c$) is locally maximal. The language $L(A_{abb}(\{u, v\})$ contains the language $aabcab(b \cdot c)$ and, in particular, the word $w = abcabbc$. The word $w' = abcabbc$ is another superword of u and v, and $w' \leq w$. Thus, w is not an locally maximal superword of u and v.
5. 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: $|LCS(u, v)| + |SCS(u, v)| = |u| + |v|$.

2.

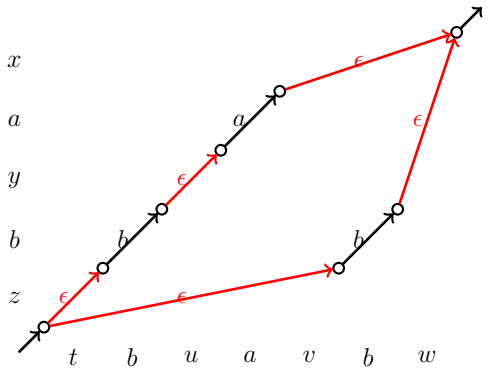
$$\begin{aligned} |W| &= |x^1| + |w_1| + \ldots + |x^p| + |w_p| + |x^{p+1}| \\ &= |w| + \sum_{i=1}^{p+1} |x^i| \;=\; |w| + \sum_{i=1}^{p+1} \left(|u^i| + |v^i|\right) \\ &= |w| + (|u| - |w|) + (|v| - |w|) \;=\; |u| + |v| - |w| \,. \end{aligned}$$

□

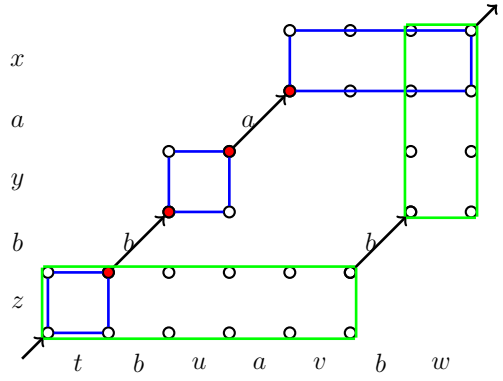
2.4. Constructive Algorithms

We have devised an algorithm producing a finite automaton $\mathcal{A}_{\sqcap(\{u,v\})}$ which exactly recognizes the language $\sqcap(\{u,v\})$, the set of locally maximal subwords common to two words u and v, due to lack of space, we do not describe it here. Its construction and proof of correctness are given in (Miclet et al., 2012). It is based on the transformation of an 2-d array displaying which letters are common to two words into a finite automaton recognizing $\sqcap(u,v)$ (see an example on figure 2(*a*)).

Starting from $\mathcal{A}_{\sqcap(\{u,v\})}$, it is then simple to produce a finite automaton that we call $\mathcal{A}_{\sqcup(\{u,v\})}$ which exactly recognizes the language $L(A(\{u,v\}))$ (also denoted $\sqcup(\{u,v\})$). We display an example at figure 2(*b*).



(*a*) An automaton which recognizes the language $\sqcap(r,s)$. We have $r = zbyax$ and $s = tbuavbw$; a and b are letters, while t,u,v,w,x,y and z are factors on $\Sigma\backslash\{a,b\}$.



(*b*) An automaton which recognizes $\sqcup(r,s) = (z \bullet t)b(u \bullet y)a(vbw \bullet x) \cup (tbuav \bullet z)b(w \bullet yax)$. A rectangle holds for the shuffle of the factors on its sides

Figure 2: Two automata built on two sentences.

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 which is *subset of* U. Moreover, we will assume that U is an antichain according to the order relation $\leq$.

3.1. Order Relation

Definition 5 (Order on alignments on U) *Given two alignments on U, $\mathtt{a} = \{(w_1, l_1), \ldots, (w_n, l_n)\}$ and $\mathtt{b} = \{(w'_1, l'_1), \ldots, (w'_m, l'_m)\}$, we write $\mathtt{a} \sqsubseteq \mathtt{b}$ 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 $\mathtt{a} \sqsubseteq \mathtt{b}$, then $word(\mathtt{a}) \subseteq word(\mathtt{b})$.

It is easy to check that $\sqsubseteq$ 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 $\mathcal{A}_H$.*

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

Definition 7 (Order on homogeneous sets of alignments on U) *Let A and B be two homogeneous sets of alignments. We have $A \sqsubseteq B$ if for all $\mathtt{b} \in B$, there is $\mathtt{a} \in A$ such that $\mathtt{a} \sqsubseteq b$.*

Proposition 2 *$\sqsubseteq$ is a partial order on $\mathcal{A}_H$ and the smallest element is $\{\emptyset\}$.*

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: $A \sqsubseteq B$ and $B \sqsubseteq A$. Since A and B are homogeneous, all alignments in A have the same support, denoted $word(A)$, and the same holds for B, with the support denoted $word(B)$. From the definition of $\sqsubseteq$, we easily check that $word(A) = word(B)$. Let us consider $\mathtt{b}_1 = \{(w_1, l'_1), \ldots, (w_n, l'_n)\} \in B$: since $A \sqsubseteq B$ and $B \sqsubseteq A$, it exists $\mathtt{a} \in A$ and $\mathtt{b}_2 \in B$ such that $\mathtt{a} \sqsubseteq \mathtt{b}_1$ and $\mathtt{b}_2 \sqsubseteq \mathtt{a}$. By transitivity, we have $\mathtt{b}_2 \sqsubseteq \mathtt{b}_1$. At last, $\mathtt{b}_1$ and $\mathtt{b}_2$ having the same support and being locally maximal, it implies that $\mathtt{b}_1 = \mathtt{b}_2$ and then $\mathtt{a} \in B$. Hence, $A \subseteq B$. Similarly, we can check that $B \subseteq A$.
□

3.2. Definition and Properties of Υ

Definition 8 *Let $\mathtt{a} \in A_r(\{u_1, \cdots, u_n\})$ and $\mathtt{b} \in A_s(\{v_1, \cdots, v_m\})$, where $\mathtt{a} = \{(u_1, l_1), \ldots, (u_n, l_n)\}$ and $\mathtt{b} = \{(v_1, l'_1), \ldots, (v_m, l'_m)\}$. Firstly, we construct $\mathtt{a} + \mathtt{b}$, the finite set of alignments $\mathtt{c} = \{(w_1, L_1), \ldots, (w_p, L_p)\}$ such that*

1. $\{w_1, \ldots, w_p\} = word(\mathtt{a}) \cup word(\mathtt{b})$
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 $\mathtt{a} \curlyvee \mathtt{b}$ *the set of minimal elements of* $\mathtt{a} + \mathtt{b}$ *according to* $\sqsubseteq$.

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

The operation $\curlyvee$ 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 $A \curlyvee B$ as the set of the minimal elements of $A + B$ according to $\sqsubseteq$ where*

$$A + B = \bigcup_{\substack{\mathtt{b} \in B \\ \mathtt{a} \in A}} (\mathtt{a} + \mathtt{b})$$

Proposition 3 *The operation $\curlyvee$ is internal to $\mathcal{A}_H$, commutative and idempotent.*

Proof. Let us consider $A \in \mathcal{A}_H$ and $B \in \mathcal{A}_H$.

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

□

3.3. Construction of $\curlywedge$

Definition 10 *Let* $\mathtt{a} \in A_r(\{u_1, \cdots, u_n\})$ *and* $\mathtt{b} \in A_s(\{v_1, \cdots, v_m\})$ *where* $\mathtt{a} = \{(u_1, l_1), \ldots, (u_n, l_n)\}$ *and* $\mathtt{b} = \{(v_1, l'_1), \ldots, (v_m, l'_m)\}$. *We construct* $\mathtt{a} \curlywedge \mathtt{b}$, *the finite set of alignments* $\mathtt{c} = \{(w_1, L_1), \ldots, (w_p, L_p)\}$ *such that*

1. $\{w_1, \ldots, w_p\} = word(\mathtt{a}) \cap word(\mathtt{b})$
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. $\mathtt{c}$ *is a locally maximal alignment.*

An alignment in $\mathtt{a} \curlywedge \mathtt{b}$ is thus based either on a restriction of $\mathtt{a}$ to the support $word(\mathtt{a}) \cap word(\mathtt{b})$ or on a restriction of $\mathtt{b}$ to the same support. For instance, if $\mathtt{a} = \{(\boxed{\mathtt{a}}cd, ab\boxed{\mathtt{a}}c, \boxed{\mathtt{a}}ba)\}$ and $\mathtt{b} = \{(a\boxed{\mathtt{c}}d, aba\boxed{\mathtt{c}}, \boxed{\mathtt{c}}a)\}$, then $\mathtt{a} \curlywedge \mathtt{b} = \{(\boxed{\mathtt{a}}\boxed{\mathtt{c}}d, \boxed{\mathtt{a}}ba\boxed{\mathtt{c}}), (\boxed{\mathtt{a}}\boxed{\mathtt{c}}d, ab\boxed{\mathtt{a}}\boxed{\mathtt{c}})\}$.

Definition 11

$$A \curlywedge B = \bigcup_{\substack{\mathtt{b} \in B \\ \mathtt{a} \in A}} (\mathtt{a} \curlywedge \mathtt{b})$$

Proposition 4 *The operation* $\curlywedge$ *is internal to* $\mathcal{A}_H$*, commutative and idempotent.*

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

□

3.4. Structure of Homogeneous Sets of Alignments on U

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

Proposition 5 *Let* A *and* B *be finite homogeneous sets of alignments. Then* $\sup_{\sqsubseteq}(A, B)$ *exists and*

$$\sup_{\sqsubseteq}(A, B) = A \curlyvee B$$

Proof.

- First, we show that $A \curlyvee B$ is greater than A and B for $\sqsubseteq$. Let $\mathtt{c} \in A \curlyvee B$. By construction, there exist $\mathtt{a} \in A$ and $\mathtt{b} \in B$ such that $\mathtt{c} \in \mathtt{a} \curlyvee \mathtt{b} \subseteq \mathtt{a} + \mathtt{b}$. By the first item of definition 8, $word(\mathtt{a}) \subseteq word(\mathtt{c})$ and by the two other items, we can conclude that $\mathtt{a} \sqsubseteq \mathtt{c}$. Thus for every $\mathtt{c} \in C$ there is $a \in A$ such that $\mathtt{a} \sqsubseteq \mathtt{c}$. Thus $A \sqsubseteq A \curlyvee B$ and $B \sqsubseteq A \curlyvee B$.
- Let C be a set of alignments greater than A and B, and let $\mathtt{c} \in C$. There are $\mathtt{a} \in A$ and $\mathtt{b} \in B$ such that $\mathtt{a} \sqsubseteq \mathtt{c}$ and $\mathtt{b} \sqsubseteq \mathtt{c}$. We need to find $\mathtt{c}' \in A \curlyvee B$ such that $\mathtt{c}' \sqsubseteq \mathtt{c}$. Remove from the support of $\mathtt{c}$ all words not in the support of $\mathtt{a}$ or $\mathtt{b}$. The obtained alignment may not be locally maximal, so we add more boxed letters to make it locally maximal. The result alignment $\mathtt{c}'$ satisfies all conditions of Definition 8, thus $A \curlyvee B \sqsubseteq C$ and therefore $\sup_{\sqsubseteq}(A, B) = A \curlyvee B$.

□

There is no equivalent relation between $\curlywedge$ and 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* $\mathcal{A}(U) = \{A(V) \mid V \subseteq U\}$.

Proposition 6 *Let* A *and* B *be sets of alignments in* $\mathcal{A}(U)$*. Then, in* $\mathcal{A}(U)$*,* $\inf_{\sqsubseteq}(A, B)$ *exists and:*

$$\inf_{\sqsubseteq}(A, B) = A \curlywedge B$$

Proof.

- First, we show that if $A = A(V)$ and $B = A(W)$ with $V \subseteq U$ and $W \subseteq U$ then $A \curlywedge B = A(W \cap V)$. Let $\mathtt{c} \in A \curlywedge B$. $\mathtt{c}$ is a loccally maximal alignment on its support $word(A) \cap word(B) = W \cap V$, thus $\mathtt{c} \in A(W \cap V)$. Let $\mathtt{c} \in A(W \cap V)$. Let $\mathtt{a}$ be an alignment on W such that $\mathtt{c} \sqsubseteq a$, then $\mathtt{c}$ is obtained from $\mathtt{a} \in A$ using the definition of $A \curlywedge B$ and $\mathtt{c} \in A \curlywedge B$.

- Let $C \in \mathcal{A}(U)$ be a set of alignments smaller than A and B. We show that C is smaller than $A \curlywedge B$. Some alignments of C are smaller than alignments of A and others are smaller than alignments of B. Since C is homogeneous, its support $word(C)$ must be included in $word(A) \cap word(B)$ and since $C = A(T)$ for some $T \subseteq U$, then $T \subseteq V \cap W$. Therefore $A(T) \sqsubseteq A(V \cap W)$ which is exactly $C \sqsubseteq A \curlywedge B$.

□

Proposition 7 *Let $U = \{u_1, u_2, \cdots, u_n\}$ be a finite set of words, the operations $\curlywedge$ and $\curlyvee$ are internal to U.*

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

□

Proposition 8 *Let $U = \{u_1, u_2, \cdots, u_n\}$ be a finite set of words, antichain for $\leq$. Then $\mathcal{U} = (\mathcal{A}(U), \curlyvee, \curlywedge)$ 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.

□

4. Lattices of Alignments and Learning by Analogy

This section intends to give some hints on how the lattice structure that we have constructed on a finite language U can be used in machine learning. Actually, the alignments in the lattice reflect in a straightforward manner what a subset of sentences share, in terms of subsequences. Hence, a direct application could be to imagine the construction of an finite or infinite language (expressed intensionnally) based on the common core of U and able to give a measure of adequation of others words of $\Sigma^\star$ to U. Firstly, we give some quick definitions on the concepts of analogical proportion and its applications to words. Then we will give a definition and a result on analogical proportions in lattices. Finally, we will give preliminary results and ideas on the application of machine learning by analogy to the extension of U according to the lattice $\mathcal{U}$.

4.1. Analogical proportion : a definition

Definition 13 (Axioms of analogical proportion) *An* analogical proportion *on a set $\mathbb{E}$ is a subset of $\mathbb{E}^4$ (hence, a quaternary relation) such that, for all 4-tuples A, B, C et D in relation in this order (denoted $A : B :: C : D$):*

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

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

4.2. Analogical Proportions Between Words

According to (Stroppa and Yvon, 2005) 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 $\Sigma^\star$, and we will come back later to its use in general lattices.

Definition 14 (Analogical proportions between words.)
$(x, y, z, t) \in \Sigma^\star$ *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* $(\beta_i)_{i\in\{1,n\}} \in \Sigma^\star$ *such that:*

$$x = \alpha_1 \ldots \alpha_n,\ \ t = \beta_1 \ldots \beta_n,\ \ y = \alpha_1\beta_2\alpha_3 \ldots \alpha_n,\ \ z = \beta_1\alpha_2\beta_3 \ldots \beta_n \qquad \textit{or}$$
$$x = \alpha_1 \ldots \alpha_n,\ \ t = \beta_1 \ldots \beta_n,\ \ y = \beta_1\alpha_2\beta_3 \ldots \alpha_n,\ \ z = \alpha_1\beta_2\alpha_3 \ldots \beta_n$$

and $\quad \forall i \in \{1, n\}, \quad \alpha_i\beta_i \neq \epsilon.$

Example. *reception* : *refection* :: *deceptive* : *defective* in an analogical proportion between words, with $n = 3$ and the factors : $\alpha_1 = re$, $\alpha_2 = cept$, $\alpha_3 = ion$, $\beta_1 = de$, $\beta_2 = fect$, $\beta_3 = ive$.

The authors have shown that this definition is conform to the axioms. There exists a second definition, which is given in (Miclet et al., 2008) with the associated algorithms, that verifies the axioms as well.

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

Example. Let $\Sigma = \{a, b, c, A, B, C\}$ an alphabet with the non trivial analogical proportions $a : b :: A : B$, $a : c :: A : C$ and $c : b :: C : B$ The following alignment shows that there is an analogical proportion in $\Sigma^\star$ between the four words $CaCA$, $CabBA$, bac and $babb$.

$$\begin{pmatrix} C & \boxed{\text{a}} & & C & A \\ C & \boxed{\text{a}} & b & B & A \\ b & \boxed{\text{a}} & & c & \\ b & \boxed{\text{a}} & b & b & \end{pmatrix}$$

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 (Hassena, 2011).

4.3. Analogical Proportion in Lattices

Using the factorization technique, the authors of (Stroppa and Yvon, 2005) have found that a general definition of an analogical proportion can be given in a lattice. Unfortunately, the definition they have given was uncomplete. We give here the complete one[6].

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

$$x = (x \wedge y) \vee (x \wedge z) \ \textit{ and } \ x = (x \vee y) \wedge (x \vee z) \qquad z = (t \wedge z) \vee (x \wedge z) \ \textit{ and } \ t = (t \vee z) \wedge (t \vee y)$$
$$y = (x \wedge y) \vee (t \wedge y) \ \textit{ and } \ y = (x \vee y) \wedge (t \vee y) \qquad t = (t \wedge z) \vee (t \wedge y) \ \textit{ and } \ z = (t \vee z) \wedge (x \vee z)$$

A simple example of proportion in a lattice is given by the following property:

6. H. Prade and L. Miclet, personnal communication.

Proposition 9 *Let y and z be two elements of a lattice. Then the following analogical proportion holds:*

$$(y \ : \ y \vee z \ :: \ y \wedge z \ : \ z)$$

4.4. Learning from $\mathcal{U}$

After having given the basis in the previous sections, we give preliminary here remarks and hints concerning some possible extensions of this work to applications, via machine learning, in connexion with analogical proportions and lattice structure.

Firstly, when investigating the connexions between locally maximal subwords, locally minimal superwords and analogical proportions, a first property is easy to show from definition 15 and proposition 1.

Proposition 10

$$\forall t \in L(A_w(\{u,v\})), \ \exists w \in \sqcap(u,v) \quad \textit{such that} \quad t \ : \ u \ :: \ v \ : \ w$$

$$\forall w \in \sqcap(u,v), \ \exists t \in L(A_w(\{u,v\})), \quad \textit{such that} \quad t \ : \ u \ :: \ v \ : \ w$$

Take $u = abcabb$ and $v = aabbc$ with the maximal subword $y = abb$. The alignment $(\boxed{a}\boxed{b}ca\boxed{b}b, \boxed{a}a\boxed{b}\boxed{b}c)$ is locally maximal. The language $L(A_{abb}(\{u,v\})$ contains the word $w = aabcabbc$. In the facing alignment one can observe the analogical proportion $w : u :: v : y$

$$\begin{pmatrix} a & \boxed{a} & \boxed{b} & c & a & \boxed{b} & b & c \\ & \boxed{a} & \boxed{b} & c & a & \boxed{b} & b & \\ a & \boxed{a} & \boxed{b} & & & \boxed{b} & & c \\ & \boxed{a} & \boxed{b} & & & \boxed{b} & & \end{pmatrix}$$

However, what we are really interested in is to find how using the lattice $\mathcal{U}$ and its analogical properties to generalize U. As a second remark, we note that any homogeneous set of alignments A in $\mathcal{U}$ represents an intensional definition of the finite language $\sqcup(A)$, also written $L(A(U))$, as defined at sections 2.3 and 2.4. We can also construct, as indicated in section 2.4, a finite automaton as an intensional representation of this language, with the syntactic analysis facility. Therefore, we have potentially at our disposal a lattice of finite automata, in connection with the lattice of subsets of U: each automaton recognizes a finite language which is a particular generalization of the associated support, itself a subset of U.

We denote hereafter $\leq$ the order relation between finite set of words derived from the subword relation $\leq$, defined by: $M \leq N$ iff $\forall m \in M, \ \exists \ n \in N$ 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 $\sqcap(K) \subset \sqcap(J)$.*

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 (de la Higuera, 2010). This could be one basis for the use of our lattice in machine learning.

Another threads to follow could be the idea of analogical closure of a finite language, as described in (Lepage, 2003) and that of analogical generation, see (Bayoudh et al., 2007). In both, a triple of words are taken in the learning sample and a fouth 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 (Kuznetsov, 2001). In addition, connections with the recent concept of *string extension*, see Kasprzik and Kötzing (2010), have been suggested to the authors, and it seem an interesting track to follow.

5. Conclusion and Bibliographical Comments

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. (Gusfield, 1997), pp 287-293 and 309, (Irving and Fraser, 1992)). 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 paper we have characterized, via the construction of a lattice of alignment sets, interesting sets of minimal superwords and maximal subwords from a set of words. We have not worked yet neither on the theoretical complexity of the construction of the lattice of alignments, nor on its practical complexity and applications. We have also given hints on using this lattice for the learning of intensional representations of languages generalizing a finite set of words.

A complexity result (sometimes misinterpreted) is given by Maier (Maier, 1978) 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. It is also true that finding the length of a shortest (longest) super(sub)sequence common to a set of k sequences is in $\mathcal{O}(m_1 \dots m_k)$ comparisons, with m_i the size of the i-th of the k sequences, hence exponential in k.

The works of Fraser and Irving (Fraser et al., 1996) have produced algorithms to find the *longest* minimal common supersequence (superword) and the *shortest* maximal common subsequence, according to the order relation $\leq$.

(Stroppa and Yvon, 2005) give a definition of an analogical proportion between words and also within lattices, that we have completed as indicated in definition 16. As far as we know, these are the only investigations about analogical proportions in the lattices.

Machine learning with the help of lattice structure is firstly investigated in (Mitchell, 1997). References on works on learning with Galois lattices can be found for example in (Kuznetsov, 2001). The cognitive aspects of reasoning and learning by analogy can be found in (Gentner et al., 1989). Methodology, algorithms and experiments in learning by analogical proportions can be found in (Miclet et al., 2008). Learning intensional expressions of langages from words is the matter of Grammatical Inference, see (de la Higuera, 2010).

The authors want to thank the anonymous reviewers for their help.

References

S. Bayoudh, H. Mouchère, L. Miclet, and E. Anquetil. Learning a classifier with very few examples: analogy based and knowledge based generation of new examples for character recognition. In *European Conference on Machine Learning, Springer LNAI 4701*, 2007.

C. de la Higuera. *Grammatical Inference*. Cambridge University Press, 2010.

C. Fraser, R. Irving, and M. Middendorf. Maximal common subsequences and minimal common supersequences. *Information and Computation*, 124:145–153, 1996.

D. Gentner, K. Holyoak, and B. Kokinov (Editors). *The Analogical Mind: Perspectives from Cognitive Science*. MIT Press, 1989.

D. Gusfield. *Algorithms on Strings, Trees, and Sequences*. Cambridge Univ. Press, 1997.

A. Ben Hassena. *Apprentissage par analogie de structures d'arbres*. PhD thesis, Université de Rennes 1, 2011.

R. Irving and C. Fraser. Two algorithms for the longest common subsequence of three (and more) strings. In *Proc. 3rd Symp. on Combinatorial Pattern Matching. Springer LCNS 644*, pages 214–229, 1992.

R. Irving and C. Fraser. Maximal common subsequences and minimal common supersequences. In *Proc. 5rd Symp. on Combinatorial Pattern Matching. Springer LCNS 807*, pages 173–183, 1994.

A. Kasprzik and T. Kötzing. String extension learning using lattices. In *LNCS 6031*, volume 6031, pages 380–391. Springer, 2010.

O. Kuznetsov. Machine learning on the basis of formal concept analysis. *Automation and Remote Control*, 62, Issue 10:1543 – 1564, 2001.

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

D. Maier. The complexity of some problems on subsequences and supersequences. *JACM*, 25:332–336, 1978.

L. Miclet, S. Bayoudh, and A. Delhay. Analogical dissimilarity: Definition, algorithms and two experiments in machine learning. *JAIR*, 32:793–824, 2008.

L. Miclet, N. Barbot, and B. Jeudy. The construction of a finite automaton recognizing exactly the locally maximal subwords common to two (and more) words. *Submitted*, 2012.

T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

N. Stroppa and F. Yvon. Analogical learning and formal proportions: Definitions and methodological issues. Technical Report ENST-2005-D004, ENST, June 2005.

R. Wagner and M. Fisher. The string-to-string correction problem. *Journal of the ACM*, 21 (1):168–173, 1974.