Analogical proportions in a lattice of sets of alignments built on the common subwords in a finite language

L. Miclet[1], N. Barbot[1], B. Jeudy[2]

[1]IRISA, Lannion & Rennes, France

[2]Université Jean Monnet, Saint-Étienne, France

ECAI 2012, SAMAI Workshop

IRISA

Roadmap

Dynamic Programming and inspiration

- Lectures on Dynamic Programming: examples, among which the Longest Common Subword $lcs(u, v)$ problem.
- For the exam, the Shortest Common Superword $SCS(u, v)$.
- HA-HA ! Notice that $|SCS(u, v)| + |lcs(u, v)| = |u| + |v|$. (N.B.: Not *that* smart: there is a simple demonstration).
- ANALOGY ! Analog to $LCM(a, b) \times GCD(a, b) = a \times b$. There is also an analogical proportion between the four words:

$$SCS(u, v) \ : \ u \ :: \ lcs(u, v) \ : \ v$$

Would there be a lattice of *SCS* and *lcs* generated by a finite language U, with nice analogical proportions ?
This is where transpiration begins.

A first glance

$$
\begin{array}{rcccccc}
u = cadb = & c & a & d & \sim & b & \sim \\
 & | & | & | & | & | & | \\
v = aebf = & \sim & a & \sim & e & b & f \\
 & | & | & | & | & | & | \\
lcs(u,v) = & \sim & a & \sim & \sim & b & \sim \\
 & | & | & | & | & | & | \\
SCS(u,v) = & c & a & d & e & b & f \\
 & & & e & d & &
\end{array}
$$

Roadmap

Basics of stringology

- Σ is an alphabet, *i.e.* a finite set of letters.
- A *word* $u \in \Sigma^\star$ 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 the letters at some (non necessarily adjacent) positions in u.

non necessarily adjacent	*subword*	*subsequence*	*sous-mot*
adjacent	*factor*	*substring*	*facteur*

The shuffle of words

- We denote $u \bullet v$ or $\bullet(u, v)$ the *shuffle* of two words.
- For example,
  $ab \bullet bc = \{abbc, (abbc), abcb, babc, bacb, bcab\}$
- The set $u \bullet v$ has at most $\binom{|u|}{|u|+|v|}$ elements.
- This operation is associative.

Order relations in $\Sigma^\star$

- In $\Sigma^\star$, there is a partial order relation order relation denoted $\leq$ defined by:

$$(u \leq v \Leftrightarrow u \text{ is a subword of } v)$$

- When u is a subword of v, v is called a *superword* of u. For example:

$$abc \leq aabbcd$$

$$abd \leq aabbcd$$

- Another (total) order relation is that on the lengths of the words.

non necessarily adjacent	*superword*	*supersequence*	*sur-mot*
adjacent	*superstring*		*sur-facteur* (?)

For two words

- w is a *common subword* to u and v when $w \leq u$ and $w \leq v$.
- w is a *maximal* common subword to u and v if there not exist any other common subword x to u and v such that $w \leq x$.

Definition

- $\sqcup\!\cdot(u, v)$ *is the set of maximal common subwords to* u *and* v
- $\sqcap\!\cdot(u, v)$ *is the set of minimal common superwords to* u *and* v

For two words

For example, ab and c are maximal common subwords to $u = cadba$ and $v = fagbhc$, while a is a non maximal common subword.

$$
\begin{array}{rccccccccc}
u = cadba = & c & \sim & a & d & \sim & b & a & \sim & \sim \\
 & | & | & | & | & | & | & | & | & | \\
v = fagbhc = & \sim & f & a & \sim & g & b & \sim & h & c \\
 & | & | & | & | & | & | & | & | & | \\
lcs(u,v) = & \sim & \sim & a & \sim & \sim & b & \sim & \sim & \sim \\
 & | & | & | & | & | & | & | & | & | \\
SCS(u,v) = & c & f & a & d & g & b & a & h & c
\end{array}
$$

For two words

For example, ab and c are maximal common subwords to $u = cadba$ and $v = fagbhc$, while a is a non maximal common subword.

$$u = cadba = \quad c \quad \sim \quad a \quad d \quad \sim \quad b \quad a \quad \sim \quad \sim$$

$$v = fagbhc = \quad \sim \quad f \quad a \quad \sim \quad g \quad b \quad \sim \quad h \quad c$$

Introduction to locally maximal subwords

a is a subword of ab, but...

$$
\begin{array}{rcccccccccc}
u = cadba = & c & a & d & b & \sim & a & \sim & \sim & \sim & \sim \\
 & | & | & | & | & | & | & | & | & | & | \\
v = fagbhc = & \sim & \sim & \sim & \sim & f & a & g & b & h & c \\
 & | & | & | & | & | & | & | & | & | & | \\
lcs(u,v) = & \sim & \sim & \sim & \sim & \sim & a & \sim & \sim & \sim & \sim \\
 & | & | & | & | & | & | & | & | & | & | \\
SCS(u,v) = & c & a & d & b & f & a & g & b & h & c
\end{array}
$$

Actually, $SCS(u,v) = (cadb \bullet f)\ a\ gbhc$

Call a (second in u, first in v) a *locally* maximal subword of u and v.

Back to *SCS* and *lcs*

Enumerating $\sqcup\!\!\cdot(u, v)$ is tricky !

- $u = bac$ and $v = dae$.
- The four words in $(d \bullet b)a(c \bullet e)$ are minimal superwords of length five.
- But u and v have four other minimal superwords of length six, included in $u \bullet v$, namely $ba(c \bullet d)ae$ and $da(e \bullet b)ac$.
- All the other elements of $u \bullet v$ are non minimal.

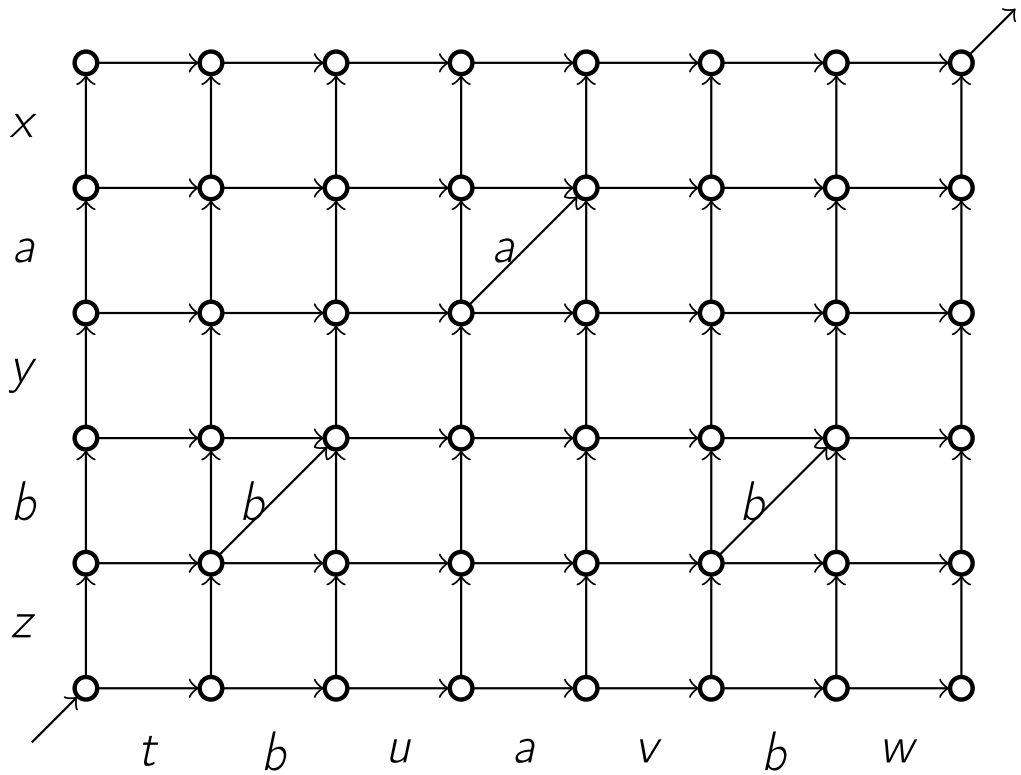
An example

- $\sqcap\!\!\cdot(cadba, fagbhc) = \{ab, c\}$
- $\sqcup\!\!\cdot(cadba, fagbhc)$ includes for example

$$cfadgbahc \quad \text{and} \quad fagbhcadba$$

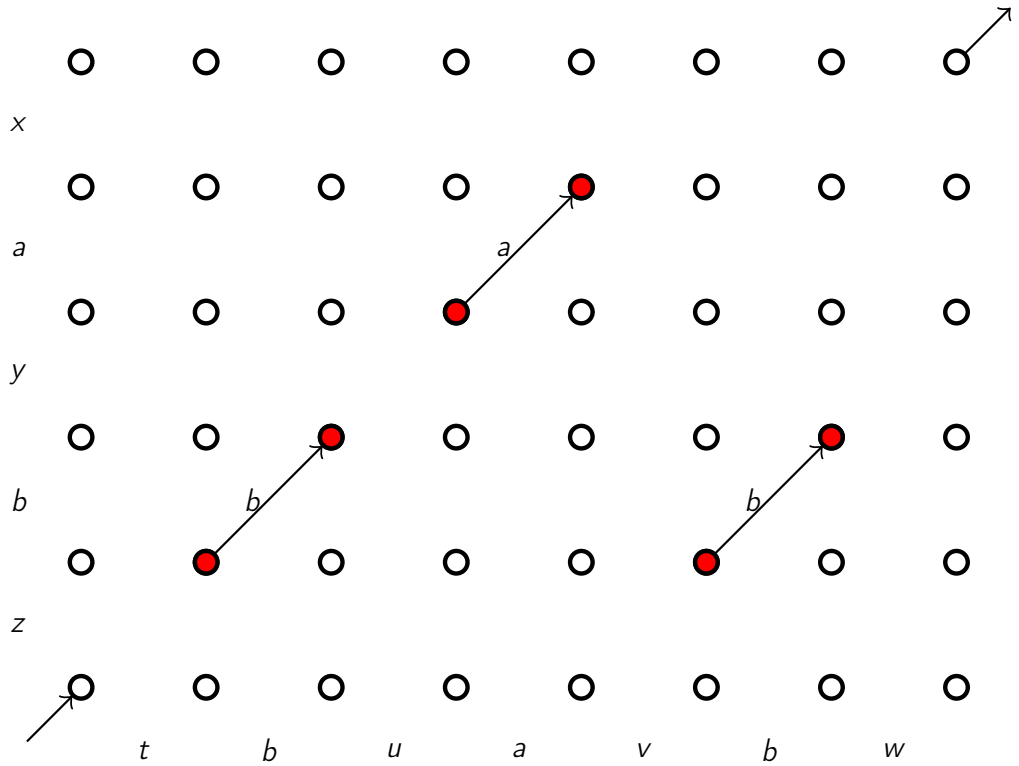
SA

Graphically speaking, a simpler example



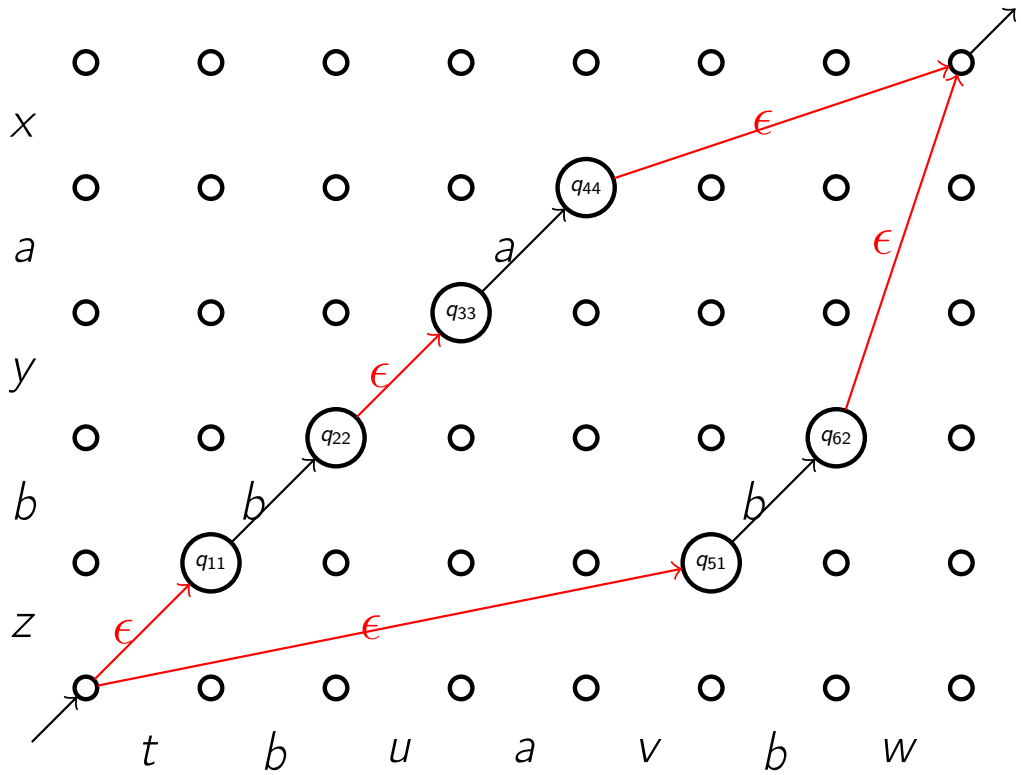
All minimal superwords of *tbuavbw* and *zbyax* (and a lot of non minimal) are recognized by this finite automaton. The maximal subword *ba* is can be read in the diagonals.

Locally maximal subwords



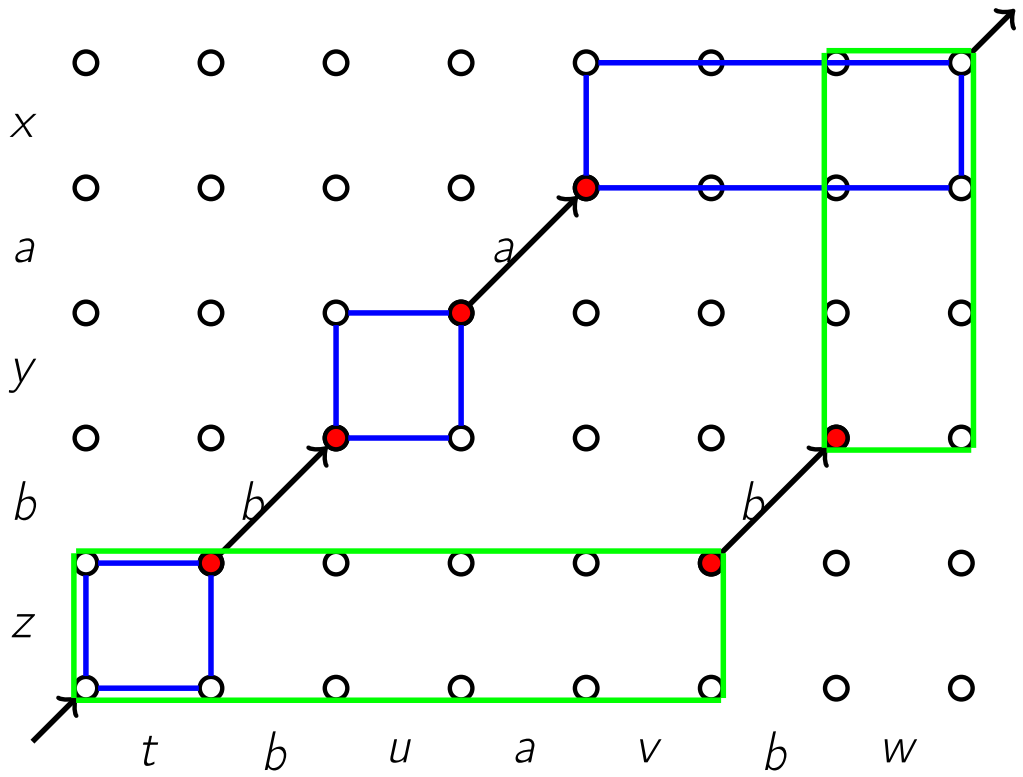
There is a partial order between the diagonal transitions.
The locally maximal subwords are the maximal chains of this order.
Construct an automaton from the Hasse diagram of this order

Locally maximal subwords and locally minimal superwords $\sqcap(u, v)$ and $\sqcup(u, v)$.



The automaton $\mathcal{A}_{\sqcap}(u, v)$ recognizing the locally maximal subwords common to u and v.

Constructing the locally minimal superwords



This automaton recognizes exactly the locally minimal superwords.
A rectangle stands for a shuffle automaton.

Conclusion of the geometric part

- Construct $\boxdot(u, v)$ (maximal subwords)?
  Construct the Hasse diagram and eliminate subwords.
- Construct $\boxplus(u, v)$ (minimal superwords)?
  Construct $u \boxdot v$ and test all subwords.
- Construct $\sqcap(u, v)$ (locally maximal subwords) ?
  Construct the Hasse diagram.
- Construct $\sqcup(u, v)$ (locally minimal superwords) ?
  By definition, from the previous.

- Generalization to any number of words: quite possible for the locally sub and superwords.
- We deal in the following only with the locally sub and superwords.

SA

Roadmap

Example: an alignment of three words on a common subword

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

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

$$L(\mathtt{a}) = (\epsilon \bullet \epsilon \bullet g)a(\epsilon \bullet \epsilon \bullet h)c(bdeg \bullet eh \bullet d)$$

Alignment

Definition

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

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

The set of words on which the alignment is defined is called the *support* and is denoted $word(\mathrm{a}) = \{w \mid \exists I \text{ with } (w, I) \in \mathrm{a}\}$.

Locally maximal alignments and subwords

Definition 1

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, $l_i \subset l'_i$.

Definition 2

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 $\boxdot(\{w_1, \ldots, w_n\})$.

Example of locally maximal alignment and subword

The set of locally maximal alignments of $W = \{ababc, cabd\}$ is $\mathcal{A}(W) =$

$$\Big\{(\boxed{a}\boxed{b}abc, c\boxed{a}\boxed{b}d), (\boxed{a}ba\boxed{b}c, c\boxed{a}\boxed{b}d),$$

$$(ab\boxed{a}\boxed{b}c, c\boxed{a}\boxed{b}d), (abab\boxed{c}, \boxed{c}abd)\Big\}$$

The set of locally maximal subwords is

$$\sqcap(W) = \{ab, c\}$$

Order relation between alignments: an example

$$\begin{pmatrix} a & \boxed{\text{b}} & \boxed{\text{c}} & \\ d & \boxed{\text{b}} & \boxed{\text{c}} & \\ f & \boxed{\text{b}} & \boxed{\text{c}} & e \end{pmatrix} \sqsubseteq \begin{pmatrix} & a & \boxed{\text{b}} & d & & \\ & a & \boxed{\text{b}} & c & & \\ d & f & \boxed{\text{b}} & a & c & e \\ b & c & \boxed{\text{b}} & & & \end{pmatrix}$$

Order relation between alignments

Given two alignments
$\mathtt{a} = \{(w_1, l_1), \ldots, (w_n, l_n)\}$ and $\mathtt{b} = \{(w'_1, l'_1), \ldots, (w'_m, l'_m)\}$, we write $\mathtt{a} \sqsubseteq \mathtt{b}$ if

- For every word w in $word(\mathtt{a})$ there exists a word w' in $word(\mathtt{b})$ such that $u \leq v$.
- and $w_i = w'_j \Rightarrow l'_j \subseteq l_i$.

Homogeneous sets of alignments and order

Homogeneous sets of alignments

A set of alignments is homogeneous when all its elements have the same support.
The family of homogeneous sets of alignments is denoted $\mathcal{A}_H$.

Order on homogeneous sets of alignments

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 $\mathtt{a} \sqsubseteq b$.

Property

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

The union operation Υ on $\mathcal{A}_H$

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 $\mathtt{a} \Upsilon \mathtt{b}$ the set of minimal elements of $\mathtt{a} + \mathtt{b}$ according to $\sqsubseteq$.

The union operation Υ on $\mathcal{A}_H$

$$\mathtt{a} = \begin{pmatrix} a & \boxed{\mathtt{b}} & \boxed{\mathtt{c}} & d & \\ & \boxed{\mathtt{b}} & \boxed{\mathtt{c}} & & a \end{pmatrix} \qquad \mathtt{b} = \begin{pmatrix} & \boxed{\mathtt{b}} & & \boxed{\mathtt{d}} & a & b \\ a & \boxed{\mathtt{b}} & c & \boxed{\mathtt{d}} & & \\ a & \boxed{\mathtt{b}} & c & \boxed{\mathtt{d}} & a & \end{pmatrix}$$

$$\mathtt{a}+\mathtt{b} = \begin{pmatrix} a & \boxed{\mathtt{b}} & c & d & & \\ & \boxed{\mathtt{b}} & c & & a & \\ & \boxed{\mathtt{b}} & & d & a & b \\ a & \boxed{\mathtt{b}} & c & d & & \\ a & \boxed{\mathtt{b}} & c & d & a & \end{pmatrix} \qquad \mathtt{a} \,\Upsilon\, \mathtt{b} = \begin{pmatrix} & \boxed{\mathtt{b}} & & d & a & b \\ a & \boxed{\mathtt{b}} & c & d & a & \end{pmatrix}$$

Note that $\mathtt{a} \sqsubseteq \mathtt{a} \,\Upsilon\, \mathtt{b}$ and $\mathtt{b} \sqsubseteq \mathtt{a} \,\Upsilon\, \mathtt{b}$.

The union operation $\curlyvee$ on $\mathcal{A}_H$

Extension

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

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

Property

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

An intersection operation

$\mathtt{a} = \{(\boxed{\mathtt{a}}cd, ab\boxed{\mathtt{a}}c, \boxed{\mathtt{a}}ba)\}$ $\qquad$ $\mathtt{b} = \{(a\boxed{\mathtt{c}}d, aba\boxed{\mathtt{c}}, \boxed{\mathtt{c}}a)\}$

- The support of $\mathtt{a} \curlywedge \mathtt{b}$ id the intersection of the supports of $\mathtt{a}$ and $\mathtt{b}$, namely $\{acd, abac\}$.
- The boxed subwords have to be maximal.

$$\mathtt{a} \curlywedge \mathtt{b} = \{(\boxed{\mathtt{a}}\boxed{\mathtt{c}}d, \boxed{\mathtt{a}}ba\boxed{\mathtt{c}}), (\boxed{\mathtt{a}}\boxed{\mathtt{c}}d, ab\boxed{\mathtt{a}}\boxed{\mathtt{c}})\}$$

IRISA

An intersection operation

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 intersection operation

Extension

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

Property

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

Finding a lattice structure to $\mathcal{A}_H$

Definition

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.

Property

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

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

SA

Finding a lattice structure to $\mathcal{A}_H$

Definition

If U is a finite collection of words.
We define the collection of sets of alignments
$\mathcal{A}(U) = \{A(V) \mid V \subseteq U\}$.

Property

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

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

Finally...

Property

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

Final result

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.

Roadmap

Conclusion

Summary

- Exploring the structure of subwords and superwords common to a finite language.
- Defining the concept of locally maximal common subwords and minimal superwords.
- Finding a lattice of alignments on a finite language.

Further work

Use the lattice structure to extend finite languages using

- Analogical proportions
- Generalization