Synchronous Regular Relations and Morphological Analysis

Christian Wurm, Younes Samih,
{cwurm,samih}@phil.uni-duesseldorf.de

Abstract

We list the major properties of some important classes of subrational relations, mostly to make them easily accessible to computational linguists. We then argue that there are good linguistic reasons for using no class smaller than the class of synchronous regular relations for morphological analysis, and good mathematical reasons for using no class which is larger.

1 Below the Rational Relations

We need not stress the importance of finite state transducers and of rational relations for computational linguistics (see Johnson [1972],Koskenniemi [1983],Kaplan and Kay [1994],Beesley and Karttunen [2003]). So we rather start with stressing the importance of sub-rational relations, that is, classes of relations properly contained in the rational relations. As is well-known in the community, rational relations are not closed under intersection. Furthermore, the equivalence and inclusion problems for rational relations are undecidable. So there are a number of arguments for not using rational relations, but rather some weaker class with more favorable decision properties. The question is: if we want to go below the rational relations, which class should we choose? In the literature, we often find the so called sequential relations; these however are quite restricted and will not be considered here. We rather focus on three classes, the strictly synchronous, the k-bounded (e.g. Roark and Sproat [2007]), and the synchronous regular relations, which are ordered by inclusion. We present their main closure properties, which are partly already known. For some reason the important class of synchronous regular relations, which has attracted a lot of attention in various fields of mathematics,[1] has to our knowledge not gained very much attention in the field of computational linguistics.[2] We argue here that 1. there are good linguistic reasons for using no class smaller than the class of synchronous regular relations; and 2. we do not know of any linguistics evidence in morphology to use the more powerful rational relations instead of synchronous regular relations.

2 Closure Properties and Decision Problems

We will consider the main closure properties for classes of relations. *Union* and *intersection* of two relations R_1, R_2 are defined in the obvious set-theoretic fashion. The *complement* of a relation R is defined wrt. two alphabets Σ, T, where $R \subseteq \Sigma^* \times T^*$, and $\overline{R} := (\Sigma^* \times T^*) - R$. The *inversion* of a word $a_1...a_n \in \Sigma^*$ is defined as $(a_1...a_n)^i := a_n...a_1$. For a relation $R \subseteq \Sigma^* \times T^*$, we put $R^i := \{(w^i, v^i) : (w, v) \in R\}$. Given two relations R_1, R_2, we define their *composition* $R_1 \circ R_2 := \{(x, z) : (x, y) \in R_1, (y, z) \in R_2\}$. Given Two relations $R_1 \subseteq \Sigma_1^* \times T_1^*, R_2 \subseteq \Sigma_2^* \times T_2^*$, we define their *concatenation* $R_1 \cdot R_2 := \{(w_1w_2, v_1v_2) : (w_1, v_1) \in R_1, (w_2, v_2) \in R_2\}$. In general, we

[1] We just mention Frougny and Sakarovitch [1993], and the research on automatic structures, see Rubin [2008]

[2] We have to mention that scholars working on the finite-state manipulation platform Vaucanson have made some efforts in using synchronous regular relations, see Lesaint [2008]

Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing, pages 35–38,
St Andrews–Sctotland, July 15–17, 2013.

say a class $\mathcal{R}$ is *closed* under a n-ary operation X, if from $R_1, ..., R_n \in \mathcal{R}$ it follows that $X(R_1, ..., R_2) \in \mathcal{R}$.

3 Three Classes and Their Inclusion Relations

We consider three classes as most interesting in this context. The first one is the class of strictly synchronous regular relations (**SSR**). For generality, we present relations of arbitrary arity. R is in **SSR** if 1. R is rational, and 2. if $(w_1, ..., w_i) \in R$, then $|w_1| = ... = |w_i|$. Secondly, a relation R is k-bounded, if 1. R is rational and 2. there is a $k \in \mathbb{N}$ such that for all $(w_1, ..., w_n) \in R$, $max\{|w_1|, ..., |w_n|\} - min\{|w_1|, ..., |w_n|\} \leq k$.[3] Call this class k-**B**. Obviously, k-**B** properly contains **SSR**. As the third class, we present the synchronous regular relations (**SR**): Put $\Sigma_\perp := \Sigma \cup \{\perp\}$, for $\perp \notin \Sigma$. The **convolution** of a tuple of strings $(w_1, ..., w_i) \in (\Sigma^*)^i$, written as $\otimes(w_1, ..., w_i)$ of length $max(\{|w_j| : 1 \leq j \leq i\})$ is defined as follows: the kth component of $\otimes(w_1, ..., w_i)$ is $\langle \sigma_1, ..., \sigma_i \rangle$, where σ_j is the k-th letter of w_j provided that $k \leq |w_j|$, and $\perp$ otherwise. The convolution of a relation $R \subseteq (\Sigma^*)^i$ is defined as $\otimes R := \{\otimes(w_1, ..., w_i) : (w_1, ..., w_i) \in R\}$. A relation $R \in (\Sigma^*)^i$ is **synchronous regular**, if there is a finite state automaton over $(\Sigma_\perp)^i$ recognizing $\otimes R$.

Informally, **SR** are the relations computed by finite state transducers which allow ϵ transitions in a component only if no other letter is to follow in this component. It is not obvious that **SR** contains k-**B**; it follows however from the following well-known synchronization lemma (see Frougny and Sakarovitch [1993]):

Lemma 1 *Assume R is an n-ary rational relation, such that there is a $k \in \mathbb{N}$, such that for all $(w_1, ..., w_n) \in R$, $max\{|w_1|, ..., |w_n|\} - min\{|w_1|, ..., |w_n|\} \leq k$. Then R is in **SR**.*

4 A Logical Characterization of SR

We can actually characterize **SR** with first order logic over the language $\mathcal{L} := (EL, pref, last_a : a \in \Sigma)$ where $EL, pref$ are binary predicates, and all $a : a \in \Sigma$ are unary predicates. We call this logic $\mathsf{FOL}(\mathcal{L})$, and interpret it in the structure $\mathfrak{S} := \langle \Sigma^*, EL, pref, a : a \in \Sigma \rangle$, where Σ^* is our universe, $a : a \in \Sigma \subseteq \Sigma^*$, and $EL, pref \subseteq \Sigma^* \times \Sigma^*$. We have $w \in a$ if and only if $w = w'a$; we have $(w, v) \in pref$ if and only if $v = wv'$, that is, w is a prefix of v; and we have $(w, v) \in EL$ if and only if $|w| = |v|$. For what is to follow, we have to assume that $|\Sigma| \geq 2$. The proof of the following theorem of Eilenberg et al. [1969] is long and complicated, so we cannot even give a sketch at this place.

Theorem 2 *Assume $M \subseteq (\Sigma^*)^i$. Then there is a $\mathsf{FOL}(\mathcal{L})$-formula $\phi(x_1, ..., x_i)$ in the free variables $x_1, ..., x_i$, such that $M := \{w_1, ..., w_i \in \Sigma^* : \mathfrak{S} \models \phi(x_1, ..., x_i)[w_1, ..., w_i]\}$, if and only if $M \in$ **SR**.*

5 Mathematical Properties

5.1 Closure Properties

That **SSR** is closed under union is obvious. Intersection follows from the fact that 1. **SR** is closed under intersection, and 2. if all pairs in R_1 and R_2 have equal length, then surely the pairs in $R_1 \cap R_2$ have equal length. It is easy to see that **SSR** is not closed under complement, as the complement of $R \in$ **SSR** in particular contains all pairs of words of different length. Moreover, **SSR** is closed under inversion, because 1. rational relations are closed under inversion, and 2. equal length is preserved; **SSR** is closed under composition and concatenation for exactly the same reason. So we have quite good closure (and decision) properties; still, **SSR** is very restrictive.

Therefore one might prefer the more powerful class k-**B**. k-**B** is obviously also closed under union, closed under intersection and not under complement, for exactly the same reason as **SSR**. Also, k-**B** is closed under composition, concatenation and inversion, again for the same reasons as **SSR**.

There is a characterization of regular relations in first order logic.[4] From this result it immediately follows that **SR** is closed under union, intersection and complement, by

[3] Note the order of quantifiers: we do not fix the k for the entire class of relations; we can choose it arbitrarily for any given relation, but then it is fixed for all of its elements.

[4] Actually, this only holds for relations over an alphabet Σ with $|\Sigma| \geq 2$; but our claims are easy to show separately for the case where $|\Sigma| = 1$.

logical connectives; moreover, by logical definability we easily obtain closure under composition: put $R_1 := \{(w_1, w_2) \in (\Sigma^*)^2 : \mathfrak{S} \models \phi(x, y)[w_1, w_2]\}$; $R_2 := \{(v_1, v_2) \in (\Sigma^*)^2 : \mathfrak{S} \models \psi(y, z)[v_1, v_2]\}$; then $R_1 \circ R_2 = \{(w_1, w_2) : \mathfrak{S} \models \exists y \phi(x, y) \wedge \psi(y, z)[w_1, w_2]\}$. We can easily show that **SR** is *not* closed under concatenation: $(a, \epsilon)^* \in \mathbf{SR}$, $(b, c)^* \in \mathbf{SR}$; but $(a, \epsilon)^* \cdot (b, c)^* \notin \mathbf{SR}$.[5] As $(b, c)^* \cdot (a, \epsilon)^*$ is regular, we also know that **SR** is *not* closed under inversion.

5.2 Decision Problems

In general, the question whether for a given characterization of a rational relation R (transducer, rational expression), we have $R = \emptyset$, is decidable. From this and the fact that **SR** is a Boolean algebra it follows that for $R_1, R_2 \in$**SR**, we can decide the questions: given characterizations of R_1, R_2, is $R_1 \subseteq R_2$, and is $R_1 = R_2$? This can be demonstrated using the standard proof for regular languages. So, we have *a fortiori* the same result for **SSR**, k-**B**. For rational relations themselves the latter problems are undecidable.

6 Natural Language Morphology Requires SR

6.1 German Compounding

So which one should we take? As there is no absolutely convincing mathematical argument, we should take a look at linguistic facts. We now present an argument for using the additional power coming with synchronous regular relations.

Compounding is a very productive morphological process in German and many other languages (Dutch, Danish, Finish, Greek etc.). It is a process whereby new words are formed by combining independent words/morphemes, where there is no restriction on the number of morphemes which can be put together to form a single new word. German compounds are strictly right-headed (Toman [1992]), that is, the morphosyntactic features of the compounds are always inherited from the rightmost morpheme. The head of the compound thus determines category, gender, and all morphosyntactic features of the whole compound. For example,the `bahn` in German `Autobahn` (highway) identifies the word as singular feminine. Due to space constraints, we cannot say much about morphological analysis in general or analysis of our particular example; we will say only as much as is needed for our formal argument, which in our view however is of general importance for computational morphology.

6.2 The Compounding Relation is Synchronous Regular

If we want to morphologically analyze a compound, in a first step, we want to transduce a sequence of compounded words $W_1...W_i$ to a sequence of representations of their morphosyntactic features $C_1...C_i$. This relation is synchronous if we use words and feature bundles as atoms. One might object that this is usually not the case, or at least depends on whether we allow complex words as atomic transitions. But mathematically, we are quite safe, as we can always form a new, finite alphabet via a bijection with finite strings over another alphabets.[6] Still, this is not satisfying, as the compound is a single word, and its morphosyntactic features are exactly the same as the one of its head. As the head is rightmost, we thus have a relation of the form $(C_1...C_i, C_i)$, mapping the entire sequence to its last element. We call this the **compounding relation**, which has to be composed with the first relation. As compounding is unbounded and consequently there is no upper bound to i, this relation is *not* in k-**B**. We now show that this relation is however in **SR**. This would be obvious if the head would be the leftmost element; for the head rightmost we need some work.

Let I be a finite set, $L_i : i \in I$ a finite set of regular languages. We say a function $f : (\Sigma \times T)^* \to (\{\epsilon\} \times \{L_i : i \in I\}) \cup (\{L_i : i \in I\} \times \{\epsilon\})$ is regular, if there is a deterministic finite state automaton $(Q, \delta, q_0, \Sigma \times T)$, where δ is extended to strings in the canonical fashion, and a finite function $g : Q \to (\{\epsilon\} \times \{L_i : i \in I\}) \cup (\{L_i : i \in I\} \times \{\epsilon\})$, such that for all $(w, v) \in (\Sigma \times T)^*$, we have $f(w, v) = g \circ \delta(w, v)$.

[5] This follows from the standard proof that rational relations are not closed under intersection, which uses exactly this relation, see Kaplan and Kay [1994].

[6] Still more technically, we would also have to ensure that the bijection defines a *code*, but we leave this aside, noting that this is satisfied in all normal cases.

Lemma 3 *A relation $R \subseteq \Sigma^* \times T^*$ is in **SR**, if and only if there is a regular function f, such that for every $(w, v) \in R$, $(w, v) = (w', v') \cdot f(w', v')$, where $|w'| = |v'|$.*[7]

So take the compounding relation $\{(C_1...C_i, C_i) : C_1...C_i$ is a well-formed compound$\}$. We simply put $f(C_1, C_i) = (\{C_1\}\backslash\overline{C}_{C_1}) \times \epsilon$, where $\overline{C}_{C_i}$ is the language of well-formed compounds ending with C_i, and $L_1\backslash L_2 := \{v : \forall w \in L_1, wv \in L_2\}$; it is well-known that regular languages are closed under this operation, so the compounding relation is synchronous regular, provided that the set of compounds itself is a regular set. This is clearly the case for the languages we considered. And even if there is a language where this is not the case, this would not be an argument in particular against using **SR**, but rather against using finite-state methods in natural language morphology in general.

7 Conclusion

We have summed up the major closure and decision properties of a number of subrational classes of relations which are currently in use. The properties we listed are mostly known, and otherwise relatively easy to obtain. We have undertaken this summarization as there does not seem to be any other literature where one could find it; and in particular in the computational linguistics literature one finds very little on closure and decision properties of subrational classes of relations.

Our main argument however is of linguistic nature: we have shown that the k-bounded (and thus strictly synchronous) relations are unable to allow for morphological analysis of a phenomenon which is as common and widespread as compounding. Synchronous regular relations on the other side are powerful enough to capture this phenomenon. We also argued that synchronous regular relations are preferable over rational relations from a purely mathematical point of view, because they form a Boolean algebra and all their decision problems are decidable.

Of course, there are many finite-state NLP applications for which **SR** is insufficient, such as inserting markup expressions in shallow parsing. Our argument was: for most of standard morphological analysis, **SR** is the smallest class which provides sufficient expressive power.[8]

[7] Actually, this lemma is sometimes even taken to be the definition of **SR**; so we omit the proof.

[8] Though there are morphological phenomena which clearly go beyond the expressive power of **SR**, such as reduplication, they seem to be quite rare; and in fact, the latter is equally problematic for finite-state morphology in general as for **SR**.

References

Kenneth R. Beesley and Lauri Karttunen. *Finite state morphology*. CSLI Publ., Stanford, Calif., 2003.

Samuel. Eilenberg, C. C. Elgot, and J. C. Shepherdson. Sets recognized by n-tape automata. *Journal of Algebra*, 13:447–464, 1969.

Christiane Frougny and Jacques Sakarovitch. Synchronized rational relations of finite and infinite words. *Theor. Comput. Sci.*, 108(1): 45–82, 1993.

C. Douglas Johnson. *Formal Aspects of Phonological Description*. Mouton, The Hague, 1972.

Ron M. Kaplan and Martin Kay. Regular Models of Phonological Rule Systems. *Computational Linguistics*, 20:331–378, 1994.

Kimmo Koskenniemi. Two–level morphology. A general computational model for word–form recognition. Technical Report 11, Department of General Linguistics, University of Helsinki, 1983.

Florian Lesaint. Synchronous relations in Vaucanson. Technical Report 0833, Laboratoire de Recherche et Développement de L'Epita, 2008.

Brian Roark and Richard William Sproat. *Computational approaches to morphology and syntax*. Oxford surveys in syntax and morphology ; 4. Oxford Univ. Press, 2007.

Sasha Rubin. Automata presenting structures: A survey of the finite string case. *Bulletin of Symbolic Logic*, 14(2):169–209, 2008.

Jindrich Toman. Compound. In W. Bright, editor, *International Encyclopedia of Linguistics*, volume 1, pages 286 – 288. Oxford Univ. Pr., 1992.