LNAI 4002

Anssi Yli-Jyrä
Lauri Karttunen
Juhani Karhumäki (Eds.)

Finite-State Methods and Natural Language Processing

5th International Workshop, FSMNLP 2005
Helsinki, Finland, September 2005
Revised Papers

Springer

Anssi Yli-Jyrä Lauri Karttunen
Juhani Karhumäki (Eds.)

Finite-State Methods and Natural Language Processing

5th International Workshop, FSMNLP 2005
Helsinki, Finland, September 1-2, 2005
Revised Papers

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Anssi Yli-Jyrä
Scientific Computing Ltd.
P.O. Box 405, 02101 Espoo, Finland
E-mail: ylijyra@csc.fi

Lauri Karttunen
Palo Alto Research Center
3333 Coyote Hill Rd, Palo Alto, CA 94304, USA
E-mail: karttunen@parc.com

Juhani Karhumäki
University of Turku
Department of Mathematics
20014 Turku, Finland
E-mail: karhumak@utu.fi

Library of Congress Control Number: 2006937535

CR Subject Classification (1998): I.2.6-7, I.2, F.1.1, F.4.2-3, F.2

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-35467-0 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-35467-3 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Preface

These proceedings contain the revised versions of the papers presented at the 5th International Workshop of Finite-State Methods and Natural Language Processing, FSMNLP 2005. The book includes also the extended abstracts of a number of poster papers and software demos accepted to this conference-like workshop.

FSMNLP 2005 was held in Helsinki, Finland, on September 1–2, 2005. The event was the fifth instance in the series of FSMNLP workshops, and the first that was arranged as a stand-alone event, with two satellite events of its own: the Two-Level Morphology Day (TWOLDAY) and a national workshop on Automata, Words and Logic (AWL). The earlier FSMNLP workshops have been mainly arranged in conjunction with a bigger event such as an ECAI, ESSLLI or EACL workshop, and this practice may still be favored in the future.

The collocation of the three events promoted a multidisciplinary atmosphere. For this reason, the focus of FSMNLP 2005 covered a variety of topics related but not restricted to finite-state methods in natural language processing.

The 24 regular papers and 7 poster papers were selected from 50 submissions to the workshop. Each submitted regular paper was evaluated by at least three Program Committee members, with the help of external referees. In addition to the submitted papers and two invited lectures, six software demos were presented. The authors of the papers and extended abstracts come from Canada, Denmark, Finland, France, Germany, India, Ireland, Israel, Japan, The Netherlands, Norway, Spain, South Africa, Sweden, Turkey, and the USA.

It is a pleasure to thank the members of the Program Committee and the external referees for reviewing the papers and maintaining the high standard of the FSMNLP workshops. Naturally, we owe many thanks to every single conference participant for his or her contributions to the conference and for making FSMNLP 2005 a successful scientific event.

FSMNLP 2005 was co-organized by the Department of General Linguistics at the University of Helsinki (host) and CSC, the Finnish IT center for science (co-ordination). We thank the members of the Steering Committees for their kind support in the early stage of the project and Antti Arppe, Sari Hyvärinen and Hanna Westerlund for helping with the local arrangements. Last but not least, we thank the conference sponsors for their financial support.

August 2005

A. Yli-Jyrä
L. Karttunen
J. Karhumäki

Organization

FSMNLP 2005 was organized by the Department of General Linguistics, University of Helsinki in cooperation with CSC, the Finnish IT center for science.

Invited Speakers

Tero Harju	University of Turku, Finland
Lauri Karttunen	Palo Alto Research Center, Stanford University, USA

Program Committee

Steven Bird	University of Melbourne, Australia
Francisco Casacuberta	Universitat Politècnica de València, Spain
Jean-Marc Champarnaud	Université de Rouen, France
Jan Daciuk	Gdansk University of Technology, Poland
Jason Eisner	Johns Hopkins University, USA
Tero Harju	University of Turku, Finland
Arvi Hurskainen	IAAS, University of Helsinki, Finland
Juhani Karhumäki, *Co-chair*	University of Turku, Finland
Lauri Karttunen, *Co-chair*	PARC and Stanford University, USA
André Kempe	Xerox Research Centre Europe, France
George Anton Kiraz	Beth Mardutho: The Syriac Institute, USA
András Kornai	Budapest Institute of Technology, Hungary
D. Terence Langendoen	University of Arizona, USA
Eric Laporte	Université de Marne-la-Vallée, France
Mike Maxwell	Linguistic Data Consortium, USA
Mark-Jan Nederhof	University of Groningen, The Netherlands
Gertjan van Noord	University of Groningen, The Netherlands
Kemal Oflazer	Sabanci University, Turkey
Jean-Eric Pin	CNRS/University Paris 7, France
James Rogers	Earlham College, USA
Giorgio Satta	University of Padua, Italy
Jacques Sakarovitch	CNRS/ENST, France
Richard Sproat	University of Illinois at Urbana-Champaign, USA
Nathan Vaillette	University of Tübingen, Germany
Atro Voutilainen	Connexor Oy, Finland
Bruce W. Watson	University of Pretoria, South Africa
Shuly Wintner	University of Haifa, Israel

Sheng Yu	University of Western Ontario, Canada
Lynette van Zijl	Stellenbosch University, South Africa

Organizing Committee

Anssi Yli-Jyrä, *Chair*	University of Helsinki and CSC, Finland
Hanna-Maria Westerlund	University of Helsinki, Finland
Sari Hyvärinen	University of Helsinki, Finland
Antti Arppe	University of Helsinki, Finland

Steering Committee I (FSMNLP Traditions)

Lauri Karttunen	PARC and Stanford University, USA
Kimmo Koskenniemi	University of Helsinki, Finland
Gertjan van Noord	University of Groningen, The Netherlands
Kemal Oflazer	Sabanci University, Turkey

Steering Committee II (Local Advisory Group)

Lauri Carlson	University of Helsinki, Finland
Tero Harju	University of Turku, Finland
Lauri Hella	University of Tampere, Finland
Arvi Hurskainen	University of Helsinki, Finland
Fred Karlsson	University of Helsinki, Finland
Krista Lagus	Helsinki University of Technology, Finland
Kerkko Luosto	University of Helsinki, Finland
Matti Nykänen	University of Helsinki, Finland

Additional Referees

Rafael C. Carrasco	Universitat d'Alacant, Spain
Loek Cleophas	Technische Universiteit Eindhoven, The Netherlands
Yvon Francois	GET/ENST and LTCI, France
Ernest Ketcha Ngassam	University of South Africa and University of Pretoria, South Africa
Ines Klimann	Universite Paris 7, France
Sylvain Lombardy	Universite Paris 7, France
David Picó-Vila	Universidad Politécnica de Valencia, Spain
Enrique Vidal	Universidad Politécnica de Valencia, Spain
Juan Miguel Vilar	Universitat Jaume I, Spain
M. Inés Torres	Universidad País Vasco, Spain
Anssi Yli-Jyrä	University of Helsinki and CSC, Finland

Sponsoring Institutions

CSC - Scientific Computing Ltd., Finland
University of Helsinki, Finland
The KIT Network, Finland
Academy of Finland
Connexor Ltd., Finland
Lingsoft Ltd., Finland

Table of Contents

Invited Lectures

Contributed Papers

Abstracts of Interactive Presentations

Abstracts of Software Demos

Characterizations of Regularity

Tero Harju

Department of Mathematics, University of Turku, Finland

Abstract. Regular languages have many different characterizations in terms of automata, congruences, semigroups *etc.* We have a look at some more recent results, obtained mostly during the last two decades, namely characterizations using morphic compositions, equality sets and well orderings.

1 Introduction

We do not intend to give a full survey on regular languages but rather a short overview of some of the topics that have surfaced during the last two decades.

Customarily regular languages are defined either as languages accepted by finite automata, represented by regular expressions, or generated by right linear grammars. The most common approach is by acceptance using deterministic finite automata, or a *DFA* for short. A DFA can be described as a 'concrete machine' with a read-only input tape from which the head of the automaton reads one square at a time from the left end to the right end. A DFA $\mathcal{A}$ can be conveniently presented as a 5-tuple

$$\mathcal{A} = (Q, A, \delta, q_0, F),$$

where Q is the set of initial states, A is the alphabet of the inputs, and the *transition function* $\delta\colon Q \times A \to Q$ describes the *action* of $\mathcal{A}$ such that $\delta(q,a) = p$ means that while reading the symbol a in state q, the automaton changes to state p and starts consuming the next input symbol. The state q_0 is the initial state of $\mathcal{A}$, and $F \subseteq Q$ is the set of its final states. The action of the automaton $\mathcal{A}$ is often written in the form $qa = p$ instead of $\delta(q,a) = p$. The transition function δ extends to words by setting $\delta(q, wa) = \delta(\delta(q,w),a)$. Thus for each word w, $\delta(q,w)$ is the state where the automaton enters when started in the state q and after exhausting w. If $w = \varepsilon$, the empty word, then $\delta(q,\varepsilon) = q$ for all states q.

More pictorially a finite automaton can be described as a directed *graph*, where nodes represent the states of the automaton and each labelled edge $q \xrightarrow{a} p$ corresponds to the transition $\delta(q,a) = p$. Then $\delta(q,w)$ is the state that is reached from q by traversing the edges labelled by the letters of w.

A language $L \subseteq A$ is *regular* if it is accepted by a DFA, $L = L(\mathcal{A})$, where

$$L(\mathcal{A}) = \{w \in A^* \mid \delta(q_0, w) \in F\}.$$

A. Yli-Jyrä, L. Karttunen, and J. Karhumäki (Eds.): FSMNLP 2005, LNAI 4002, pp. 1–8, 2006.

The family of regular languages is a neat family in the sense that it is closed under many natural operations of languages: if L and K are regular languages, then so are

- $L \cup K$, $L \cap K$, $L \setminus K$, catenation $L \cdot K$, Kleene closures L^* and L^+, shuffle $K\text{s}L$, quotients $L^{-1}K$ and LK^{-1}, complement $A^* \setminus L$, morphic (and the inverse morphic) images $h(L)$ (and $h^{-1}(L)$), as well as the reversal L^R (mirror image).

This list could be continued much further.

Instead of deterministic finite automata one can also employ other finite models of automata for regular languages. For instance, a language L is regular if it is accepted by a *nondeterministic* FA where the transitions are given by a relation instead of a function.

We can extend the transition function (or relation) in several ways, say by attaching conditions to the transitions that change the design of the states. As an example, each state can have a sign, + or −, and the transitions can depend on the signs and change them.

Also, one can expand the way how finite automata accept words. An *alternating* finite automaton is a nondeterministic FA where the states are divided into existential and universal states, and acceptance depends on the global tree of behaviour.

By allowing finite automata to read the input word both to the left and right does not influence the family of accepted languages, i.e., a 2-way FA accepts only regular languages.

Decision problems for regular languages are, as a rule, decidable. However, many algorithmic problems are hard for them. For instance, one can prove that the problem of finding a nondeterministic finite automaton with the smallest number of states accepting a regular language L is truly hard. The problem is PSPACE-complete.

The syntactic characterizations of regular languages are originally due to Myhill [1] and Nerode [2] as well as to Rabin and Scott [3] at the end of the 1950s. These characterizations follow from analyzing the behaviour and structures of finite automata.

For a language $L \subseteq A^*$ define the relation $\sim_L$ by

$$u \sim_L v \iff u^{-1}L = v^{-1}L,$$

where $u^{-1}L = \{w \mid uw \in L\}$. This relation is an equivalence relation on A^*, and thus A^* is divided into equivalence classes w.r.t. $\sim_L$.

Theorem 1. *A language L is regular if and only if $\sim_L$ is of finite index, i.e., there are only finitely many equivalence classes w.r.t. $\sim_L$.*

The idea behind Theorem 1 is that the set $u^{-1}L$ corresponds to the state $\delta(q_0, u)$ of the DFA accepting L. As an example, consider the language $L = \{a^n b^n \mid n \geq 0\}$ which is well known to be nonregular. We notice that the sets $u_i^{-1}L$ are all different for the words $u_i = a^i$, $i \geq 1$. Since there are infinitely many sets $u^{-1}L$, we deduce that, indeed, the language L is not regular.

Let

$$u \cong_L v : \quad xuy \in L \iff xvy \in L$$

be the *syntactic congruence* of $L \subseteq A^*$.

Theorem 2. *A language L is regular if and only if the syntactic congruence of L has finite index.*

Using syntactic congruences one can study the fine structure of regular languages more deeply. This approach leads to *algebraic theory* of regular languages. For instance, Schützenberger [4] showed that a language L is star-free if and only if its syntactic monoid is aperiodic, i.e., contains only trivial subgroups. Here we say that L is *star-free* if it can be represented by a generalized regular expression allowing complementation L^c but disallowing stars $*$. For instance, $A^* = \emptyset^c$, and

$$(ab)^* = 1 + a\emptyset^c \cap \emptyset^c b \cap (\emptyset^c aa \emptyset^c)^c \cap (\emptyset^c bb \emptyset^c)^c.$$

We also state an algebraic characterization of regular languages that is related to syntactic congruences.

Theorem 3. *A language L is regular if and only if it is recognized by a finite monoid M, i.e., there is a finite monoid M such that $F \subseteq M$ and*

$$L = \varphi^{-1}(F)$$

for a monoid morphism $\varphi\colon A^* \to M$ onto M.

We can restate this theorem as follows:

Theorem 4. *A language L is regular if and only if there exists a finite monoid M such that*

$$L = \varphi^{-1}\varphi(L)$$

for a monoid morphism $\varphi\colon A^* \to M$.

Regular languages can also be described by matrices. The following theorem is due to Schützenberger.

Theorem 5. *For each regular language L, there are $0,1$-vectors u and v, and a matrix M (of finite sets) such that*

$$L = u^T M^* v.$$

Regular languages have had connections to logic since the studied made by Büchi [6], Elgot [7], and McNaughton and Papert [5].

Theorem 6. *A language L is regular if and only if L definable in the monadic second order logic (which allows comparisons of positions of letters in words and quantifiers over sets of positions).*

2 Morphic Characterizations

The topic of morphic characterizations of regular languages was was initiated by Culik, Fich, and Salomaa [8] in 1982, and it was continued by several people during the following years.

Recall that a mapping $h\colon A^* \to B^*$ is a *morphism* if

$$h(uv) = h(u)h(v)$$

for all words u, v. The *inverse morphism* is the many-valued mapping

$$h^{-1}(v) = \{u \mid h(u) = v\}\,.$$

In the theorems that follow the morphisms h_i, for $i = 1, 2, \ldots$, are between suitable alphabets. Culik, Fich, and Salomaa [8] proved that

Theorem 7. *A language L is regular if and only if there are morphisms h_i such that*

$$L = h_4 h_3^{-1} h_2 h_1^{-1}(a^*b).$$

This result was improved by Latteux and Leguy[9] in 1983:

Theorem 8. *A language L is regular if and only if there are morphisms h_i such that*

$$L = h_3 h_2^{-1} h_1(a^*b).$$

We shall sketch the idea behind the proof of this theorem.

In the other direction the claim follows from the fact that regular languages are closed under taking morphic images and inverse morphic images, and the starting language a^*b in Theorem 8 is certainly regular.

Let then L be a regular language and let $\mathcal{A}$ be a DFA accepting L. Assume that the states of $\mathcal{A}$ are

$$Q = \{q_0, q_1, \ldots, q_m\},$$

where q_0 is the initial state. Let

$$\Gamma = \{[q_i, x, q_j] \mid \delta(q_i, x) = q_j\}$$

be an alphabet that encodes the transitions of $\mathcal{A}$, and let a, b and d be three special symbols. Define our first morphism $h_1\colon \{a, b\} \to \{a, b, d\}^*$ by

$$h_1(a) = ad^m \quad \text{and} \quad h_1(b) = bd^m\,,$$

Hence $h_1(a^n b) = (ad^m)^n \cdot bd^m$ for each power n.

Let then $h_2\colon \Gamma^* \to \{a, b, d\}^*$ be defined by

$$h_2([q_i, x, q_j]) = \begin{cases} d^i a d^{m-j} & \text{if } j \neq m\,, \\ d^i b d^m & \text{if } j = m\,. \end{cases}$$

Hence

$$u \in h_2^{-1}h_1(a^n b) \iff u \text{ codes the accepting computation of } \mathcal{A} \text{ of } a_1a_2\dots a_n.$$

Finally, let $h_3\colon \Gamma^* \to A^*$ be defined by

$$h_3([q,x,p]) = x.$$

Then $L(A) = h_3h_2^{-1}h_1(a^*b)$.

Even a simpler variant was shown to hold by Latteux and Leguy [9]:

Theorem 9. *A language L is regular if and only if there are morphisms h_i such that*

$$L = h_3^{-1}h_2h_1^{-1}(b).$$

The special case of regular star languages has especially appealing characterization.

Theorem 10. *For any language L, the language L^* is regular if and only if there exists a (uniform) morphism h and a finite set F of words such that*

$$L^* = h^{-1}(F^*).$$

The morphic characterizations of regular languages extend partly to transductions, i.e., to many-valued mappings $\tau\colon A^* \to B^*$ computed by finite transducers. The following is due to Turakainen [10], Karhumäki and Linna [11].

Theorem 11. *Let R be a given regular language. Then for all languages L,*

$$L \cap R = h_3h_2^{-1}h_1\mu(L),$$

where $\mu\colon A^* \to A^*d$ is a marking defined by $\mu(w) = wd$ for a special symbol d.

Latteux, Leguy, and Turakainen [9, 12] showed

Theorem 12. *Each rational transductions has the forms*

$$h_4h_3^{-1}h_2h_1^{-1}\mu \quad \text{and} \quad h_4^{-1}h_3h_2^{-1}h_1\mu,$$

where μ is a marking.

The following theorem of Harju and Kleijn [13] shows that there is no algorithm to decide whether the marking μ is needed.

Theorem 13. *Iy is undecidable whether or not a transduction has a representation without endmarker μ.*

3 Equality Sets

In the Post Correspondence Problem, *PCP* for short, the problem instances are pairs (g, h) of morphisms $g, h\colon A^* \to B^*$, and the problem asks to determine whether there exists a nonempty word w such that $g(w) = h(w)$. It was shown by Post in 1947 that the PCP is undecidable in general, that is, there does not exist an algorithm for its solution.

The set of all solutions of an instance $g, h\colon A^* \to B^*$ is called the *equality set* of g and h. It is the set

$$E(g,h) = \{w \in A^* \mid g(w) = h(w)\}.$$

Choffrut and Karhumäki [14] have shown that the equality set $E(g, h)$ is regular for a special class of morphisms, called bounded delay morphisms. However, for these morphisms the problem whether or not $E(h, g)$ contains a nonempty word remains undecidable! This means that there is no effective construction of a finite automaton $\mathcal{A}$ accepting the regular language $E(g, h)$ when the instance g, h consisting of bounded delay morphisms is given.

A morphism $h\colon A^* \to B^*$ is called a *prefix morphism*, if for all different letters $a, b \in A$, the image $h(a)$ is not a prefix of the image $h(b)$.

If A and B are alphabets such that $A \subseteq B$, then the morphism $\pi_A\colon B^* \to A^*$, defined by

$$\pi_A(a) = \begin{cases} a & \text{if } a \in A, \\ \varepsilon & \text{if } a \in B \setminus A, \end{cases}$$

is the *projection* of B^* onto A^*.

The next result is due to Halava, Harju, and Latteux [15, 16].

Theorem 14. *A star language $L = L^* \subseteq A^*$ is regular if and only if*

$$L = \pi_A(E(g,h))$$

for prefix morphisms g, h and the projection π_A onto A^*.

A morphism $f\colon A^* \to B^*$ is a *coding*, if it maps letters to letters.

Theorem 15. *A star language $L = L^* \subseteq A^*$ is regular if and only if*

$$L = f(E(g,h))$$

for prefix morphisms g, h and a coding f.

4 Well Quasi-orders

A *quasi-order* $\rho \subseteq X \times X$ on a set X is a reflexive and transitive order relation:

$$x\rho x \quad \text{and} \quad \left.\begin{matrix} x\rho y \\ y\rho z \end{matrix}\right\} \Longrightarrow x\rho z.$$

Moreover, ρ is a *well quasi-order*, *wqo* for short, if every nonempty subset $Y \subseteq X$ has at least one minimal element but only finite number of (non-equivalent) minimal elements. In the below instead of ρ we use $\leq$ for an order relation.