

Optimalitáselmélet és véges állapotú módszerek

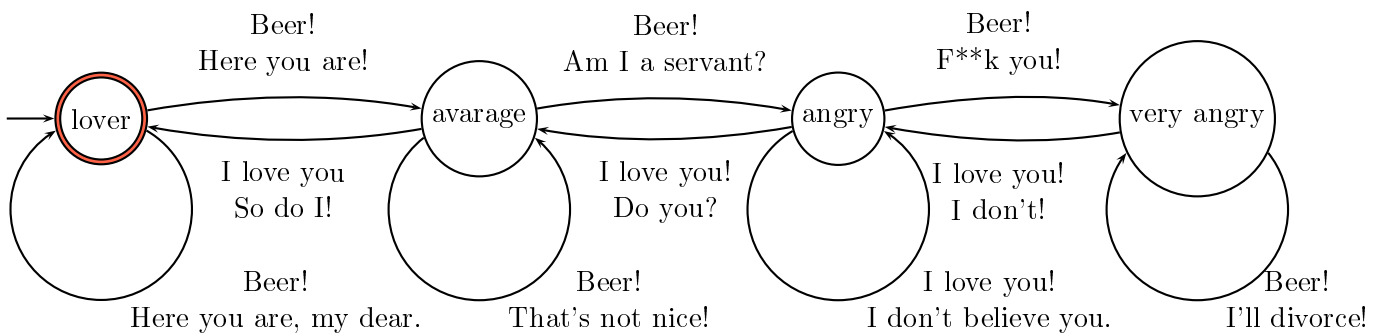
Tamás Bíró

2003. október 14.

1 Véges állapotú automaták / transzducerek

Véges állapotú automaták által elfogadott nyelvek = reguláris grammatikával leírható nyelvek
 = reguláris kifejezéssel megadható nyelvek.

Hogyan képzelik el a férfiak a nőket?



E szerint a modell szerint az alábbi párbeszéd grammatikus:

$$\left\{ \begin{array}{l} \text{Beer!} \\ \text{Here you are!} \end{array} \right\} \left\{ \begin{array}{l} \text{Beer!} \\ \text{Am I a servant?} \end{array} \right\} \left\{ \begin{array}{l} \text{I love you!} \\ \text{Do you?} \end{array} \right\} \left\{ \begin{array}{l} \text{Beer!} \\ \text{That's not nice} \end{array} \right\} \left\{ \begin{array}{l} \text{Beer!} \\ \text{That's not nice} \end{array} \right\} \left\{ \begin{array}{l} \text{I love you} \\ \text{So do I!} \end{array} \right\}$$

Ez viszont agrammatikus:

$$* \left\{ \begin{array}{l} \text{Beer!} \\ \text{Here you are!} \end{array} \right\} \left\{ \begin{array}{l} \text{I love you!} \\ \text{Do you?} \end{array} \right\} \left\{ \begin{array}{l} \text{I love you!} \\ \text{I don't!} \end{array} \right\}$$

Miért nem működik egy párkapcsolat, ha egy férfi így gondolkodik?

Nincs hosszútávú memória a modellben!

Hogyan lehet ilyen modellt komoly dolgokra is használni?

Memória hiánya miatt nem alkalmazható például a következők leírására:

- long-distance dependencies (Chomsky 1957: az angol szintaxis nem írható le reguláris nyelvként)
- reduplikatív morfológia (kiv. ha véges lexikon,...)

2 Véges állapotú Optimalitáselmélet

Johnson, 1972: az SPE környezetfüggő szabályai többsége valójában reguláris.
Három lehetséges architektúra a véges állapotú morfo/fonológiában:

- Párhuzamos (two-level morphology, Kimmo Koskeniemi, 1983.)
- Szeriális (Kaplan and Kay, 1994.)
- OT (Frank and Satta, 1998; Karttunen, 1998; Gerdemann and van Noord, 2000; Jäger, 2002.)

Optimalitáselmélet:

- Gen: leképezés az UR-ról a candidate-halmazra
- Eval: szintről szintre kiszűrni a nem optimálisakat:
 - * Megfogalmazni a constraint-et
 - * szűrés, az összes vetélytárs függvényében

3 A Gen, mint véges állapotú leképezés

Van, ami nem (biztos, hogy) megvalósítható, pl. reduplikatív morfológia.
Szótagszerkezet: ld. Karttunen, 1998.; Gerdemann and van Noord, 2000.
Metrikus hangsúly:

$$word = \# \left| \left\{ \begin{array}{l} unparsed\ syllable \\ non-head-foot \end{array} \right\}^* \right| head-foot \left| \left\{ \begin{array}{l} unparsed\ syllable \\ non-head-foot \end{array} \right\}^* \right| \#$$

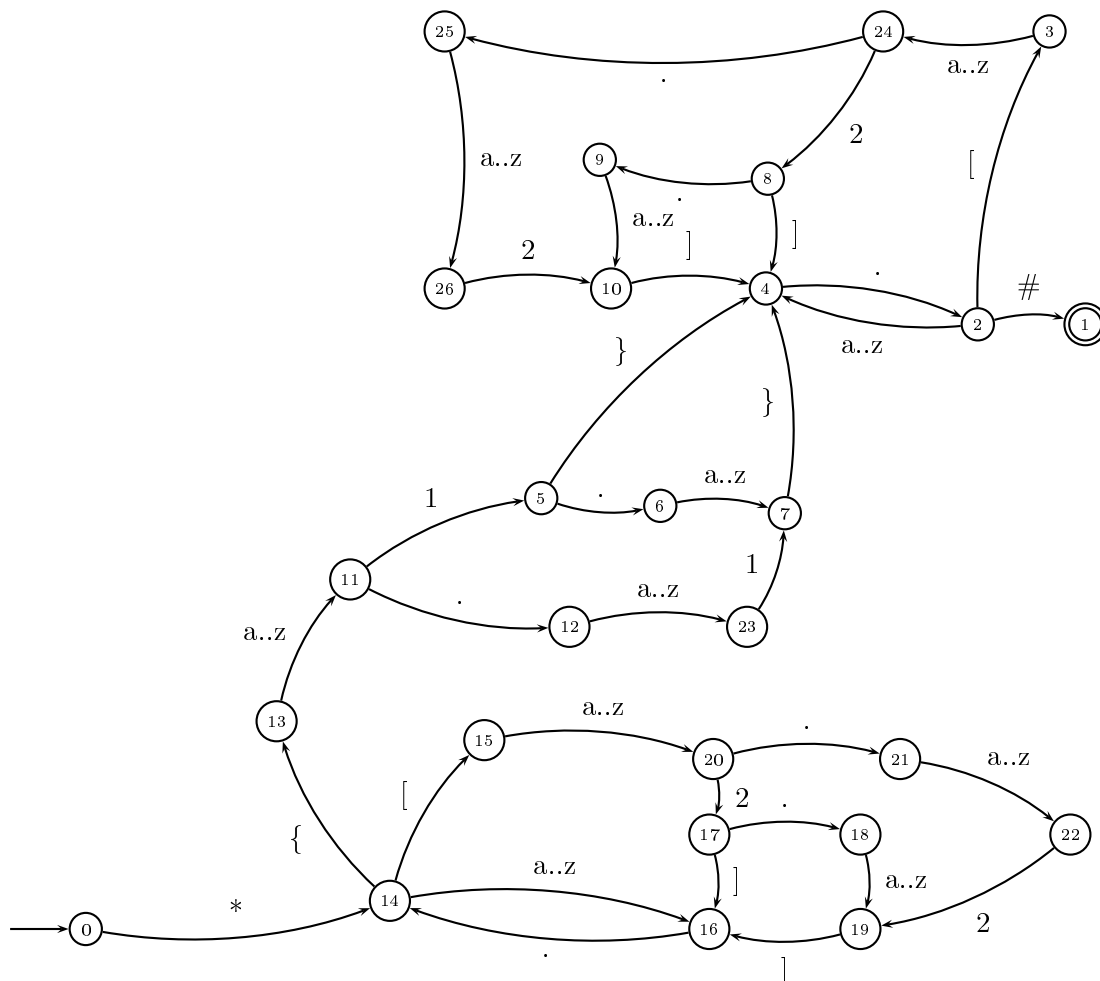
$$unparsed\ syllable = phonemes^*|.$$

$$non-head-foot = \left\{ \begin{array}{l} phonemes^*|2|. \\ phonemes^*|2|.phonemes^*|. \\ phonemes^*|.phonemes^*|2|. \end{array} \right\}$$

$$head-foot = \left\{ \begin{array}{l} phonemes^*|1|. \\ phonemes^*|1|.phonemes^*|. \\ phonemes^*|.phonemes^*|1|. \end{array} \right\}$$

Ezek reguláris kifejezések, amelyekből egyszerűen megépíthető egy transzducer (pl. Gertjan van Noord *FSA Tools*-a: <http://www.let.rug.nl/~vannoord/Fsa/>).

Az eredmény:



* ab.ra.ka.dab.ra.#



FST in state S

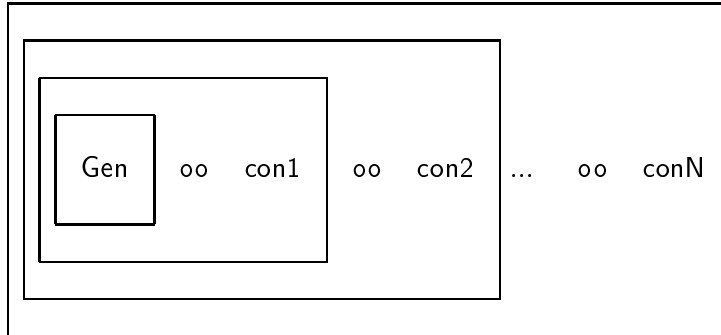


* ab.{ra.ka1}.[dab2.ra].#

4 Az Eval megvalósíthatósága a FS OT-ban

Probléma: az, hogy adott jelölt túlél-e egy szintet, függ attól, hogy milyen más jelöltek kerültek még be az adott szintbe.

$$\text{Inp } oo \text{ con}_i := \{ \langle i, o \rangle \mid \langle i, o \rangle \in \text{Inp} \text{ és } \forall \langle i, o' \rangle \in \text{Inp} : \langle i, o \rangle \succeq \text{con}_i \langle i, o' \rangle \}$$



Két feladatunk van:

- Megfogalmazni az egyes constraint-eket FST-ekkel.
- Megoldani a szűrést.

Frank and Satta (1998.) és Karttunen (1998.):

Lenient composition:

- Két szint (van sértés vagy nincs sértés).
- Minden constraint-hez egy transducer: a constraint-et kielégítő sztringek halmaza:

$$T_C := Ident_{\{w \mid w \in \Sigma^*, C(w)=0\}}$$

- Szűrés:

$$OT \circ C := (OT \circ T_C) \cup (Ident_{\overline{domain(OT \circ T_C)}} \circ OT)$$

- Véges n számú szint: szűrők sorozata (counting approach).

Gerdemann and van Noord (2000.):

Matching approach:

- Nincs határ a sértések számára.
- Minden constraint-hez FST (T_C): a sértéseket beírjuk a sztringbe.
 - Bizonyos mértékű lokalitás
 - Max. lineáris constraintek
- Kreáljuk meg a nem-optimális jelöltek halmazát:
 - További sértés-jelek hozzáadása (adviol).
 - Sértés-jelek permutálása n -szer (permute_mark _{n}).
 - Ha túl sok sértés: ez csupán n -ed rendű közelítés.

- Szűrés:

$$OT \circ C := (OT \circ T_C) \circ Ident_{\overline{range(OT \circ T_C \circ addviol \circ permute_mark_n)}} \circ delete_mark$$

Jäger (2002.):

Generalized matching approach:

- Nincs határ a sértések számára.
- T_C : leképez minden jelöltet a nála kevésbé optimális jelöltek halmazára.
- Szűrés:

$$OT \circ C := OT \circ \text{Ident}_{\overline{\text{range}(OT \circ T_C)}}$$

5 Hangsúly-constraintek FS OT-vel

A constraint-ek javasolt tipológiája:

1. Maximálisan 1 (vagy konstans) sértés / szó.
2. A sértések maximális száma arányos a szó hosszával.
3. A sértések maximális száma gyorsabban nő, mint a szó hossza.

1. típus: a szűrés könnyen megvalósítható (Frank and Satta 1998, Karttunen 1998):

- ALIGN(Word, Foot, Left): align left edge of word with left edge of some foot.

Nem biztos, hogy a constraint megfogalmazható FS-ként (természetes nyelveknél előfordulhat?
Ld. J. Eisner-féle OTP):

- MATCHESOUTPUTOFSPE: The output matches the result of applying Chomsky & Halle (1968) to the input. (J. Eisner, 1999)

2a. típus: a sértésjelek száma max. arányos a szó hosszával, és a sértés-jelek “szépen” helyezkednek el (pl. a szó széléhez igazodnak).

- ALIGN(Main-foot, Word, Left): align head-foot with word, left edge.
 $\sigma * \sigma * \sigma * [\sigma \sigma] \sigma \sigma$

Az előforduló constraint-ek megfogalmazhatóak, és a szűrés megvalósítható (Gerdemann és van Noord (2000), 0 permutációval).

2b. típus: a sértésjelek max. száma arányos a szóhosszal, de bárhol elhelyezkedhetnek.

- Parse-syllable: each syllable must be footed.
- Iambic: align the right edge of each foot with its head syllable.

A szűrés nem megvalósítható! (Kivéve, ha maximalizáljuk a sértések számát, azaz a szó hosszát, akkor n permutációval.)

A használt constraint-ek többsége ide tartozik: lokálisak, és a sértésjeleket kihelyező T_C megfogalmazható.

3. típus: a sértésjelek száma gyorsabban nő, mint a szó hossza. Pl. kvadratikusan constraint-ek (v.ö. Eisner (1997.), Bíró (2003.)):

- ALIGN(Foot,Word,Left): align each foot with the word, left edge.

Már a sértésjeleket se tudjuk kiosztani (a 'pumping lemma' következménye):

Theorem: Let \mathbf{T} be a functional finite state transducer. Then there exists a linear upper bound on the length of the output, *i.e.* there exists a positive integer N such that for any input string σ (for which there exists an output $T(\sigma)$) the following holds:

$$|T(\sigma)| \leq N |\sigma|$$

Tanulság:

Hypotheses underlying OT (explicit in McCarthy 2002):

- *Locus hypothesis:* A violation mark is assigned for each *locus* of violation within a candidate.
- *Gradience hypothesis:* Some constraints are gradient: multiple violations to a single locus.
- *Homogeneity hypothesis:* Multiple violations of a constraint from either source are added together in evaluating a candidate.

ALIGN(Main-foot,Word,Left): gradient, de átfogalmazható.

Ne használjunk "gradient constraint"-eket: fogalmazzuk át, vagy kerüljük ki őket:

- McCarthy (2002.) szerint "they are harmful"
- Erősebb generatív erejük van, mint FS.

6 Tanulás FS OT-val

7 Irodalom

Douglas C. Johnson (1972), *Formal Aspects of Phonological Description*

Robert Frank and Giorgio Satta (1998): *Optimality Theory and the Generative Complexity of Constraint Violability*, Computational Linguistics 24, 2, pp. 307-315.

Lauri Karttunen (1998): *The proper treatment of optimality theory in computational phonology*, in: *Finite-state Methods in NLP*, pp. 1-12, Ankara.

Jason Eisner: *Doing OT in a Straitjacket*, ld. J. Eisner honlapján.

Dale Gerdemann and Gertjan van Noord (2000): *Approximation and Exactness in Finite State Optimality Theory*, in: Jason Eisner, Lauri Karttunen, Alain Thériault (eds): *SIGPHON 2000, Finite State Phonology*.

Gerhard Jäger (2002): *Gradient constraints in finite state OT: The unidirectional and the bidirectional case*, ROA-479; későbbi változat: *Recursion by optimization: On the complexity of bidirectional Optimality Theory*, Natural Language Engineering.

John J. McCarthy (2002): *Against Gradience*, ROA-510.

Tamás Bíró: *Quadratic Alignment Constraints and Finite State Optimality Theory*, Proceedings of the Workshop on Finite-State Methods in Natural Language Processing (FSMNLP), in the framework of EACL 10, Budapest.