651.926.011

A GÉPI /MONDAT/ANALIZIS NÉHÁNY KÉRDÉSÉRŐL

Ábrahám Sámuel

Napjaink egyik legfontosabb tudományága az i n f o r m á c i ó f e l d o l g o z á s é s v i s s z a n y e r é s elmélete és gyakorlata. Modern tudományunk és technikánk mind nagyobb és nagyobb mennyiségü információ feldolgozását és visszanyerését igényli éspedig meghatározott, legtöbbször rövid időn belül. Ezeknek az igényeknek a kielégitésére az emberi erő egyedül már nem elégséges. Ezért égető sürgősséggel merül fel az információ feldolgozás és visszanyerés g é p e s i t é s é n e k problémája, vagyis különböző gépek felhasználása az információ feldolgozás és visszanyerés folyamatában. A folyamatra engedélyezett idő határozza meg, hogy mechanikus vagy elektronikus gépek használata a legcélszerübb, és egyáltalán nem kell arra törekedni, hogy minden egyes esetben függetlenül az adott körülményektől, mindent a leggyorsabb elektronikus gépekre bizzunk. Mindamellett, hogy elméletileg az egész folyamat egységes gépesitése tünik optimálisnak, gyakorlatilag egy megfelelő ember-gép rendszer kidolgozása /ez leginkább talán a gépi forditás esetében látszik/ a legkedvezőbb megoldás.

Mivelhogy a mai társadalomban az információ legnagyobb része természetes nyelvek segitségével rögzitődik, az információ feldolgozás és visszanyerés /egyik/ legfontosabb része a természetes nyelveken irt /ritkán: kiejtett/ szövegek feldolgozása. Ezért a jelen tanulmányban csak a természetes nyelveken irt szövegek g é p i feldolgozásával foglalkozunk, pontosabban az ilyen szövegekből kapott információ gépi visszanyerésének egyes kérdéseivel. A kapcsolatban levő mondatok / = szövegek/ gépi feldolgozásának kérdései csak a legutóbbi időben kerültek az érdeklődés középpontjába, mert az egyszerübb kérdés, a mondatok gépi feldolgozása sem nyert még elégséges megoldást.

A mondatok gépi feldolgozása elméletileg egyenlő a mondatok formális feldolgozásának fogalmával. A formális feldolgozást már pontosan meg lehet határozni: egy feldolgozás formális, ha csak a mondatot alkotó jelek formáját és mondaton belüli sorrendjét veszi tekintetbe /ezt és csak ezt tudja a gép érzékelni/. Hangsulyoznunk kell, hogy formális feldolgozás révén /mindamellett, hogy határai mindinkább tágulnak/ lehetetlen a mondatban tartalmazott teljes információt visszanyerni. /A következőkben a "mondat formális feldolgozása a benne tartalmazandó információ visszanyerésének céljából"

kifejezés helyett a "mondat formális analizise" kifejezést használjuk, és amikor ez nem vezethet félreértéshez, a "formális" szót elhagyjuk./

Az utóbbi idők kutatásai arra utalnak, hogy a formális analizis könnyebben megvalósitható, ha először egy olyan /formális/ rendszert épitünk fel, amely előállitja /generálja/ a mondatokat, és utólag "forditjuk" meg ezt a rendszert. Ezért észlelhetjük napjainkban a generativ rendszerek /generativ grammatikák/ iránti nagy érdeklődést.

Ezeknek a grammatikáknak a szerkesztését és tanulmányozását N.CHOMSKY /1/ inditotta el, /és folytatja számottevő eredménnyel./ CHOMSKY szerint egy g e n e r a t i v m o d e l l /most már világos, hogy a "grammatika"szó használata nem teljesen indokolt/ egy olyan formális utasitás-rendszer /device/, amely az adott nyelv összes mondatait /és csak azokat/ állitja elő, mégpedig ugy, hogy a generálás miként je alapján már egyértelműen meghatározhatjuk a /generált/ mondat összes formális tulajdonságait. Ezt ugy is mondhatjuk, hogy a generálás folyamatában minden mondathoz ez az utasitás-rendszer adekvát analizist kapcsoljon. Természetesen a fenti elgondolás csak akkor válik ténylegesen meghatározássá, ha pontosan megfogalmazzuk, mit értünk "formális utasitás-rendszer"-en, és melyek a mondat "összes" formális tulajdonságai. Az elsőt megtehetjük. Formális utasitás-rendszer alatt egy $G = /V, \Sigma, F/$ hármast értünk, ahol V egy véges szótár, amely véges számu szimbólumot tartalmaz. Σ egy részhalmaza V-nek. F pedig egy véges halmaz, amely $/\varphi, \psi/$ ketteseket tartalmaz, ahol φ, ψ a V-ben tartalmazott szimbólumok sorozatai. Az F-be tartozó ketteseket törvényeknek nevezzük. Egy ilyen utasitás-rendszer a következőképpen generál:

Azt mondjuk, hogy σ_i -re alkalmazható a f_k / = $/\varphi_k, \psi_k//$ törvény, ha /a/ σ_i olyan szimbólumok sorozata, amelyek V elemei, /b/ σ_i tartalmazza φ_k és /c/ teljesül egy adott C_k feltétel. Ha /a/, /b/, /c/ teljesül, akkor f_k /egyszeri/ alkalmazása σ_i -re abból áll, hogy leirjuk a σ_j szimbólum sorozatot, amely σ_i -től csak annyiban különbözik, hogy σ_i -ban /az első/ φ_k helyett ψ_k -t irtunk. Ebben az esetben azt mondjuk, hogy σ_i /közvetlenül/ generálja σ_j -t. Terminális derivációnak egy olyan $\sigma_1 \ldots \sigma_n$ sorozatot nevezünk, amelyben /a/ σ_i generálja σ_{i+1} -et /minden $1 \leq i \leq n-1$ -re/, /b/ σ_i eleme Σ-nak és /c/ σ_n nem generál semmit. σ_n -t mondatnak, és a mondatok halmazát, L_G-t, a G utasitás-rendszer által generált nyelvnek nevezzük.

Egyszerü példája a magyar nyelv kis részét generáló modellnek a következő:

G_m : V = {S, NP, VP, A, N, az, a, ember, könyvet, felveszi}
= {S}

F = /1/ /S, NP VP/
/2/ /NP, A N/
/3/ /VP, V NP/
/4/ /A, {az, a} /C_4/
/5/ /N, {ember, könyvet} /C_5/
/6/ /V, felveszi/.

Mielőtt valamit is generálnánk ezzel a modellel, két megjegyzést kell tennünk. A { } zárójel azt jelenti, hogy az egyik vagy a másik szimbólumot választjuk ki, a C_4 /vagy C_5/-ben megfogalmazott feltételek szerint. A feltételeket /C_4, C_5/ célszerű a törvények leirásánál hozzájuk kapcsolni. /Az /1/, /2/, /3/, /6/ törvények alkalmazása nincs semmilyen feltételhez kötve./

Ha C_4 és C_5-öt megfelelően fogalmazzuk meg, akkor G_m a következőképpen generál egy magyar mondatot /két szimbólum-sorozat közé irjuk az alkalmazott törvény számát/.

S /1/, NP VP /2/, A N VP /3/, A N V NP / /,

A N V A N /4/, Az N V A N /5/, az ember

V A N /6/, az ember felveszi A N /5/,

az ember felveszi a könyvet /4/, a z e m b e r f e l v e s z i a k ö n y v e t .

Sokkal nehezebb a feladatunk, ha az "összes formális tulajdonságok" fogalmát akarjuk pontosan meghatározni. Általában ezt a fogalmat csak magyarázzák, az adott nyelvet beszélő egyének "nyelvi intuiciójával", /amely fogalom ugyancsak nincs meghatározva/.

Az F-ben tartalmazott törvények és a hozzájuk kapcsolt C_k feltételek formájának különféle meghatározásával különböző generativ modelleket kapunk. Közülük érdemlegesek az ugynevezett

/1/ Kontextus mentes mondatszerkezet modellje,

/2/ Kontextust tekintetbe vevő mondatszerkezet modellje,

/3/ Transzformációs modell. /2/

Meg kell jegyeznünk, hogy mig az /1/ és /2/ modellek esetében sikerült pontosan meghatározni a törvények és feltételek formáját, a transzformációs modell esetében ezt még kielégitő módon elérni nem sikerült.

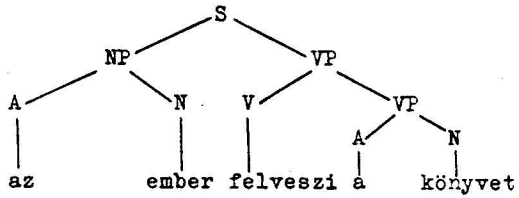
A jelen tanulmány értelmében részünkre különösen fontos az,hogy a modell ne csak generálja az adott természetes nyelvet, /vagyis generativ ereje legyen/, hanem hogy a generálás folyamata minden egyes mondathoz egy megfelelő analizist is kapcsoljon /vagyis explikativ ereje is legyen/.

Az /1/ és /2/-es modellekben minden generált mondathoz egy cimkével ellátott ágrajzot /labelled-tree, P-marker/ kapcsolunk ugy, hogy a terminális derivació tagjait sorrendben egymás alá irjuk és vonallal kötjük össze azokat a szimbólumokat /két egymásután következő sorban/, amelyek együtt egy f_k törvényt alkotnak.

A fenti példa esetében,

a z e m b e r f e l v e s z i a k ö n y v e t

- mondatnak a következő P-marker felel meg



CHOMSKY-nak sikerült kimutatni, hogy az /1/ és /2/ modellek /a fenti megfogalmazásokban/ még ha generálják is az adott természetes nyelvet /ez eddig nem sikerült ezt a kérdést teljesen tisztázni/ az általuk biztositott analizis nem adekvát. Ezért az egyetlen helyes modellnek a transzformációs modellt tekintik, amelynek egy része egy /2/ tipusu modell. Ebben a modellben először véges számu un. m a g m o n d a t o t /pontosabban majdnem mondatot/ generálnak, amelyekből kiindulva generálják az összes többi mondatokat. Ebben a modellben egy mondat analizise, ha magmondatról van szó, egy cimkével ellátott ágrajz, ha nem magmondatról van szó, akkor több cimkével ellátott ágrajz /amelyek között vannak a generálásnál felhasznált magmondatok cimkével ellátott ágrajzai/.

Ideje, hogy áttérjünk a fenti modellek "megforditása" lehetőségeinek tanulmányozására. Az /1/ és /2/ modelleket meg lehet fogalmazni mint fél-asszociativ-rendszerek, és ezen az alapon minden egyes ilyen modellnek egyértelmüen meg lehet feleltetni egy Markov-féle normál algoritmust. /3/ A Markov-féle normál algoritmusokat pedig könnyü "megforditani": Minden egyes $\mathcal{A}$ Markov-féle normál algoritmus esetében fel lehet épiteni /bizonyos feltételek mellett egyértelmüen/ egy olyan $\mathcal{A}$ Markov-féle normál algoritmust, hogy ha $\mathcal{A}$/P/ = R, akkor $\mathcal{A}$/R/ = P. A probléma részletezését l.: /4/ és /5/-ben. Mivel a Markov-féle normál algoritmusok gépileg megvalósithatók, a fenti modellek esetében a gépi analizis teljes egészében elvégezhető.

Amint már emlitettem, az /1/ és /2/ tipusu modellek, CHOMSKY megfogalmazásában, nem megfelelőek a természetes nyelvek leirására, a transzformációs modell, /amelyről feltételezhetjük, hogy megfelelő/ pedig sajnos még nincs formális szempontból kielégitően megfogalmazva, igy "megforditásuk" kérdését még csak fel sem lehet vetni. Ebből a helyzetből három kiut lehetséges.

CHOMSKY és követői azon fáradoznak, hogy a transzformációs modell egy formális szempontból kielégitő meghatározását adják. Ezek fontos, és szerintünk eredménnyel kecsegtető munkálatok, de eddig végleges eredmények még nem születtek.

A kutatók egy kisebb csoportja ujabb fajta modelleket próbál felépiteni, amelyek eredményesebbek lesznek mint az /1/ és /2/-es tipusnál. /Az eddigi eredmények viszonya a transzformációs grammatikákhoz nincs tisztázva./ Korai lenne még eredményekről beszélni.

Szerintünk egy harmadik ut is járható. Tul erősnek tünik

CHOMSKY-nak az a követelménye, hogy egy generativ modell ugy legyen felépitve, hogy már a generálás alapján adja ki a megfelelő analizist. Ha ezt a két követelményt, a generálás és a megfelelő analizis felépitésének követelményét, szétválasztjuk, akkor az előttünk álló feladat megoldhatóvá válik. Eddig sikerült bebizonyitani, hogy ha adva van egy transzformációs grammatika /az eddigi megfogalmazások értelmében/, akkor fel lehet épiteni egy vele generativ ekvivalens, vagyis egy pontosan ugyanazt a nyelvet generáló, /2/-es tipusu grammatikát.

Ebben a grammatikában ujabb formális müveletek bevezetésével minden egyes, már generált mondathoz egy olyan analizist lehet kapcsolni, amely egyenértékü a transzformációs grammatikában kapott analizissel. Ezt az uj modellt //2/-es tipusu grammatika + uj analizis rendszer/ már meg lehet fogalmazni mint Markov-féle normál algoritmust, és ebből következően meg lehet "forditani".

Az ilyen ujfajta, általunk generativ-analitikus-nak nevezett modellek elméletén és gyakorlati kérdésein KIEFER Ferenc tudományos munkatárssal együtt dolgozunk a Magyar Tudományos Akadémia Számitástechnikai Központja keretében, és az eredményekről a közeljövőben részletesen beszámolunk.

BIBLIOGRÁFIA

/1/ CHOMSKY,N.: Syntactic Structures. The Hague, 1963. Monton and Co. 118 p.

/2/ KIEFER F.: Matematikai nyelvészet. Bp. 1964. OMKDK 179 p.

/3/ MARKOV,A.A.: Teoria algorifmov. Trudi. mat. inst. AN SSR, XLII, Moszkva, 1964. AN SSR 375 p.

/4/ ÁBRAHÁM,S.: A formal study of generative grammars 1, Computational linguistics /Budapest/ 1964, II.sz. 5-21.p.

/5/ ÁBRAHÁM,S.: A formal study of generative grammars 2. Computational linguistics /Budapest/ 1965. IV.sz. /nyomás alatt/

oOOo

Ш. Абрахам
Некоторые вопросы машинного анализа /предложений/

В статье рассматриваются известные до сих пор методы формального анализа /графически изображенных/ предложений в рамках порождающих грамматик. После указания недостатков этих методов, показываются, что они вытекают из /слишком сильного/ требования, чтобы анализ был полностью и исключительноо обоснован на процессе порождения предложения. Если от этого требования отказаться, то порождающую грамматику можно дополнить системой определений таким образом, чтобы удовлетворительный формальный анализ стал полностью осуществим. Такая работа проводится в рамках Вычислительного центра ВАН. Более подробно новый метод будет изложен в будущей статье.

ÁBRAHÁM, S.: Some questions of the mechanical analysis of sentences

In the present paper methods of formal language analysis /within generative grammars/ are presented. After discussing their shortcomings, it is shown that latter are due to the /to strong/ restriction that the analysis should be based solely on the way the sentence has been generated. If we do not respect this restriction, generative grammar can be completed by a conceptual machinery so that adequate formal analysis becomes possible even on the base of lower-type generative grammars. Studies on such systems are carried out at the Computing Centre of the Hungarian Academy of Sciences, and will be reported in a future paper in more detail.