

## A KORSZERŰ DOKUMENTÁCIÓ Néhány kérdése

Ábrahám Sámuel

Senki előtt sem kétséges, hogy a gyors és eredményes tudományos és műszaki fejlődés egyik legfontosabb feltétele a meglévő /mind bel-  
földi, mind külföldi/ tudományos és műszaki információk gyors felhasználása.

A tudományos és műszaki információk gyors felhasználásának problémája egy csomó új kérdést vetett fel, amelynek a tanulmányozását és megoldását nem lehetett az ismert tudományágakon belül megvalósítani. Így született meg egy új tudományág: a z i n f o r m á c i ó t á r o l á s á n a k é s v i s s z a n y e r é s é n e k az elmélete, majd gyakorlata. Ma már mindenki elismeri, hogy ez az új tudomány egyike századunk legfontosabb vívmányainak. Pontosabban, századunk egyik legfontosabb vívmánya lesz, mert az információ tárolásának és visszanyerésének elmélete még gyerekcipőben jár. Mind a mai napig még arról is folynak viták, hogy mi is ennek az elméletnek és gyakorlatnak a tárgya. Vannak, akik már ki is fejtették negatív álláspontjukat az információ tárolás és visszanyerés gépi megoldásának lehetőségével szemben. Szerencsére ezeknek a száma nagyon kicsi és annak ellenére, hogy már évek óta kardoskodnak álláspontjuk mellett, mind az elméleti, mind a gyakorlati munka továbbfolyik, és már szép /habár nem végleges/ eredményeket is fel tud mutatni.

A fent elmondottak szerint kell megvizsgálni az információ tárolásának és visszanyerésének célkitűzéseit, az azokhoz vezető lehetséges utakat és az ebből adódó gyakorlati következtetéseket.

Azt a gyakorlati szükségletet, amely az információ tárolásának és visszanyerésének elméletéhez vezetett, a következő példával lehet /nem pontosan/ megvilágítani: tegyük fel, hogy minden egyes hazai és külföldi könyvet és folyóiratot a fővárosban egy üres házban tárolunk úgy, hogy érkezési sorrendben lerakjuk őket a soronlévő üres szobában vagy szobareszben. És most ugyancsak tételezzük fel, hogy valakit érdekel, mit irtak eddig egy bizonyos kérdésről /pl. az antibiotikumokról/, és ezzel a kérdéssel a raktárház vezetőjéhez fordul. Mit tehet a vezető, hogy a kérdésnek megfelelően, természetesen minél rövidebb időn belül?

Természetesnek tűnik a következő válasz: el kell olvasnia az

egész meglévő anyagot, eközben megjegyeznie, utána pedig átadnia a vonatkozó irodalmat a rendelőnek. Ez a megoldás nemcsak természetes, hanem az egyetlen elméletileg teljes megoldása a felvetett kérdésnek. Egyetlen hibája csak az, hogy erre talán egy egész emberélet sem volna elegendő. Ezért nem a probléma teljes megoldására kell törekedni, hanem egy olyan "legjobb" megoldásra, amelynél egy optimális időn belül kapott eredményben a szóbanlévő területre /antibiotikumok/ vonatkozó, és meg nem talált anyag lehetséges száma kisebb egy előre meghatározott számnál. Ezen kívül még annak a feltételnek is eleget kell tennünk, hogy a kérdés megoldására szükséges munka gazdaságossági szempontból ne haladjon túl egy adott határt. A megoldásra adott határidő és a kötött gazdaságossági határ fontossá teszi annak a kérdésnek a megítélését is, hogy mennyiben lehetséges elméletileg és mennyire érdemes gyakorlatilag a gépek alkalmazása a kérdés megoldására. Már itt megjegyezzük, hogy gépek alatt nemcsak elektronikus számológépeket értünk, hanem lyukkártya és egyéb berendezéseket is.

A felvetett probléma megoldása kétféle műveletet igényel:

1. A beérkező anyag /könyvek és folyóiratok/ olyan elrendezése, feldolgozása és nyilvántartása, hogy az adott és hasonló feladatok minél könnyebben legyenek megoldhatók és

2. az előzetesen rendezett, feldolgozott és nyilvántartott anyag olyan feldolgozása, amely megadja a kapott feladat megoldását.

Ennek megfelelően az információ tárolásának és visszanyerésének elmélete és gyakorlata is két részből áll:

I. előkészítő folyamat és

II. üzemeltetési folyamat.

Az előkészítő folyamat maga is több részből áll. Az általános vélemény az, hogy a befutott anyagot először is tárgy szerint csoportosítják. /Általában valamilyen tizedes rendszerű osztályozás alapján./ A valóságban az osztályozás /kategorizálás/ már egy több részből álló műveletfolyamatnak a végeredménye /vagy végcélja/. A cél elérésének alapvető feltétele, hogy a befutott anyagot megfelelően /1.1/ indexeljük /subject-heading/, vagyis röviden, néhány szóval megadjuk, hogy mi is a könyv vagy cikk tárgya. Ezt a műveletet napjainkban ember végzi, miután elolvasta a szóbanforgó könyvet vagy cikket. Miután a befutott anyagot tárgyszavazzuk, nem marad más hátra, mint hogy az adott tizedes osztályozásban kikeressük a megfelelő jelzeteket. A tapasztalatok szerint eléggé kétséges, hogy két külön dolgozó ember egy adott könyv tartalmát azonos tárgyszavakkal fejezi ki. Még kétséesebb, hogy az, aki valamilyen tárgyra vonatkozó irodalmat keres, követni fogja-e a szakozók gondolatmenetét a feladat megfogalmazásánál.

A befutott anyag előzetes feldolgozásának fontos része az is, hogy minden könyvnek vagy cikknek elkészítjük a rövid kivonát /1.2/. A kivonat felépítésének az a célja, hogy egy adott kérdés esetében magának a kérőnek ne kelljen az egész könyvet vagy cik-

ket elolvasni, hanem csak egy /sokkal/ rövidebb szöveget, amelyből már megállapítható legyen az, hogy ténylegesen szükséges-e végigolvasni az eredeti teljes anyagot.

Elvi szempontból a tárgyszavazás és a kivonatolás között nincs különbség: a tárgyszó minimális kivonat.

Ha a feldolgozásra váró anyag idegen nyelvű, akkor természetesen szükséges annak előzetes lefordítása /1.0/, pontosabban, a tárgyszó és kivonat szerkesztőjének ismernie kell az adott nyelvet.

Ami az üzemeltetési folyamatot illeti, az két részből áll: /2.1/ A kapott információ-kérés megfelelő feldolgozása, ami abból áll, hogy a kérést a tárgyszavak terminológiájába fogalmazzák át; és /2.2/ azoknak az anyagoknak a kikeresése, amelyek a /2.1/-ben kapott tárgyszavak témaköréit tartalmazzák.

Mint ahogy már említettük, a kért anyag kikeresésére adott idő és a megoldás kötött gazdaságossági határa arra készítette és készíteti a szakértőket, hogy megpróbálják a fenti kérdések megoldásánál a gép felhasználását is.

A kivonatolás alapján véve abból áll, hogy /1<sup>o</sup>/ fel kell ismerni az adott szövegben egy bizonyos tartalmi egységet és ezután /2<sup>o</sup>/ határozatot kell hozni - bizonyos összehasonlítási feltételek alapján - minden egyes talált tartalmi egység esetében.

Mivelhogy a könyvek és cikkek valamilyen természetes nyelven vannak írva, a kivonatolás feladata egy nyelvi anyag feldolgozását jelenti. Ennek a gépi megoldása ugyancsak egy új tudományághoz vezetett, a nyelvi anyag feldolgozásának /language-data processing/ elméletéhez és gyakorlatához. Nyelvi anyag-feldolgozással van dolgunk nemcsak a kivonatolásnál, hanem a fordításnál és a beérkezett információkérés megfelelő feldolgozásánál is.

A gépi kivonatolás természetesen azzal kezdődik, hogy a kivonatolásra váró anyagot bevezetjük a gépbe. Ezzel a céllal a szöveget lyukkártyákra, lyukszalagokra vagy mágneses szalagokra veszik fel. Így kell eljárni az információ kéréssel is. Ez a művelet azonban túlságosan költséges és elég hosszadalmas is. Ezért már évek óta folynak /biztató eredményekkel/ az automata beviteli /automatic input/ megvalósítására törekvő kísérletek. Ezeknek a kutatásoknak az az első célja, hogy megvalósítsák az írás és beszéd automata felismerését /automatic character and speech recognition/.

A nyelvi anyag gépi feldolgozása feltételezi az adott szöveg grammatikai felismerését is. A gépi fordítás esetében ehhez a fordításban szereplő két nyelv ismert nyelvtanainak alapján látunk hozzá. Ezen az úton azonban nem haladhatunk abban az esetben, amikor egy ismeretlen nyelvű szövegről van szó, vagy amikor a gépi feldolgozásnak célja éppen az adott szöveg nyelvi szerkezetének leírása /ami annyit jelent, hogy a szövegben lévő alapelemeket keressük - ez a szöveg szegmensálása, és az így talált alapele-

mek klasszifikálása ezeknek a releváns együttes előfordulás tulajdonságai alapján - ez a s z ö v e g d i s z t r i b u c i o n á l l i s a n a l i z i s e /. Ennek a célnak az elérésére N.D.ANDREJEV, neves leningrádi tudós statisztikai modellje látszik a legalkalmasabbnak. Ennek a modellnek ismertetését a közeljövőben fogják közreadni egy kiadványban, amely ANDREJEV nemrég Magyarországon tartott előadásait tartalmazza.

Ugyancsak a gépi kivonatolást igyekeznek megkönnyíteni /vagy lehetővé tenni/ a Z.S.HARRIS amerikai professzor /University of Pennsylvania/ által vezetett munkálatok is, amelyeknek az a céljuk, hogy a természetes nyelven írt mondatokat egy ún. magmondatokból álló szöveggé alakítsák át, amely magmondatok a legmesszebbmenőben megközelítenek egy bizonyos kívánt szerkezetet.

Az adott szövegben bizonyos tartalmi egységek felismerése a leg-egyszerűbben a thezaurus-módszerrel látszik elérhetőnek. Ez abból áll, hogy megadunk egy bizonyos szójegyzéket és egy bizonyos mondatjegyzéket, és azoknak jelenlétét vizsgáljuk meg az adott szövegben.

Ez a módszer első pillantásra utopisztikusnak tűnik, de a gyakorlatban eléggé alkalmazható. Például a fizika szótára kb. 10 000 szóból áll. Ebből a 10 000 szóból sajnos kb. 100 billió két és háromszavas mondatot lehet alkotni. Ha ellenben kizárjuk a grammatikailag vagy szemantikailag helytelen mondatokat, és csak azokat vesszük tekintetbe a megmaradt mondatok közül, amelyek többször találhatók az irodalomban, akkor már csak egy néhány ezer mondat marad. Azért, hogy össze tudjuk hasonlítani a talált tartalmi egységeket és határozni is tudjunk, minden thezaurus-elemet súlyozunk és mindig a legnehezebb elemeket tartjuk meg /egy bizonyos alsó határig, amely változhat/.

Ezenkívül fontosak azok a kutatások is, amelyeknek céljuk, hogy az adott szöveg szinonimáit /lexikai, morfológiai és szintaktikai értelemben/ jelezze ill. kizárja, azaz jelezze és kizárja a logikailag ismétlődő részeket.

A nyelvi anyag gépi feldolgozása nem szorítkozik csak a fentiekre. Az ilyen feldolgozás nagy segítségére van a nyelvész-kutatónak mind lexikai kutatásaiban, mind különböző statisztikai munkák elvégzésénél /stilus kutatás/, és a nyelvtanítás egyes kérdéseinek megoldásánál is.

Abban az esetben, amikor a feldolgozásra váró anyag idegen nyelvű, előzetes fordítás szükséges. Ez nem jelenti azt, hogy minden befutott idegen nyelvű könyvet vagy cikket le kell fordítani. Gyakorlatilag csak azt jelenti, hogy a könyvet /vagy cikket/ feldolgozás céljából egy az adott nyelvet ismerő feldolgozónak kell elolvasnia, azután készíti el és fordítja le a kivonatot. Az ilyen munkafolyamatot /a gép alkalmazásával/ a következőképpen vázolhatjuk fel:

| magyar nyelvű<br>anyag                | orosz nyelvű<br>anyag | angol nyelvű<br>anyag    |
|---------------------------------------|-----------------------|--------------------------|
| g /magyarul/                          | p i /oroszul/         | kivonatolás<br>/angolul/ |
| g é p i f o r d i t á s               |                       |                          |
| m a g y a r n y e l v ű k i v o n a t |                       |                          |

Az a tény, hogy nem magát a beérkezett anyagot fordítjuk, hanem csak a kivonatot, erősen lecsökkenti a /gépi/ fordítás volumenét.

A gépi fordítás elvi kérdéseivel szerző részletesen foglalkozik a "Computational Linguistics" 3-as számában megjelenő cikkében. Itt csak egynéhány megjegyzést közlünk.

A gépi fordítás kérdései hamarabb vonták magukra a kutatók figyelmét, mint a gépi információ tárolás és visszanyerés egyéb kérdései. Az utóbbi tíz év eredményei ellenben arra mutatnak, hogy mégis nagyon távol vagyunk ennek a kérdésnek gyakorlati megoldásától /nem beszélve a kérdés gazdaságossági oldaláról/. Ez a helyzet egyeseket arra késztetett, hogy a gépi fordítás elvi lehetetlenségének álláspontjára helyezkedjenek /180° fordulattal a saját előbbi álláspontjukkal szemben/. Eddig két /de tekintélyes/ híve van ennek az álláspontnak: Y. BAR-HILLEL és A. OETTINGER.

OETTINGER érvelése nem tartalmaz semmi újabb bizonyítékot BAR-HILLEL-éhez képest /mindamellett, hogy érint egy néhány új érdekes elvi kérdést./ BAR-HILLEL bizonyítéka pedig egész teljességében az ő általa adott meghatározáson alapszik: egy gépi fordítás akkor Teljesen Automatikus Kitűnő Minőségű Gépi Fordítás /röv.: TAKMGF, angolul FAHQMT/, ha mindennemű utólagos emberi beavatkozás nélkül egy magasképzettségű, kétnyelvű emberi fordító által kiadott fordítással azonos.

Arra a kérdésre, hogy ez valóban helyes meghatározása-e a gépi fordítás célkitűzéseinek, választ kapunk, ha alaposabban szemügyre vesszük, milyen információ alapján fordít egy magas képzettségű emberi fordító. Az emberi fordító a fordításnál kétféle információt használ fel:

1. Az adott szövegben tartalmazott információt és
2. az adott szövegen kívül álló információt.

Az adott szövegben tartalmazott információ is többféle:

1.1 nyelvi információ

2.2 nem nyelvi információ.

Napjainkban sajnos még a szövegben tartalmazott nyelvi információt sem tudjuk teljesen kihámozni, de ennek eddig legalább is nem látszik semmi elvi akadály.

A nem nyelvi információ alatt azt az információt értjük, amely érzelmeinken keresztül vitatódik át az egyik nyelvből a másikba. Például egy vers fordításánál a fordító sok esetben a nyelvi információt többféleképpen is átadhatja a fordítás nyelvén. Ilyen esetekben azokat a szavakat és konstrukciókat választja ki, amelyek benne ugyanazokat az érzelmeket váltják ki, mint az eredeti szavak és konstrukciók. Természetesen egy ilyen nem nyelvi információt a gép elvileg nem tud kivonni az adott szövegből, tehát átvinni sem tudja azt. Szerencsére ez csak a szépirodalmi fordítás esetében baj, ahol az ilyen nem nyelvi információknak határozó szerepük van /különböző irodalomfajokban, különböző fokon/. A tudományos irodalom a szépirodalommal szemben csak elenyésző százalékban /vagy egyáltalán nem/ tartalmaz ilyen fajta nem nyelvi információt. Ezért az, hogy a gép nem tudja felhasználni az efajta információt, nem érinti a tudományos irodalom gépi fordításának elvi lehetőségét.

Egy adott szövegen kívülálló információ, amelyet egy magas képzettségű emberi fordító felhasználhat, ugyancsak kétféle:

2.1 Szakismeret,

2.2 Általános ismeretek.

A szakismeretet be lehet programozni a gépbe /de ez nem gazdaságos/. Az általános /emberi/ ismereteket végtelen jellegűknél fogva, elvileg nem lehet beprogramozni a gépbe. És ebben van igazsága Y. BAR-HILLEL-nek: ebből a szempontból egy gép soha sem vetélkedhet egy emberrel. De vizsgáljuk meg, milyen hasznunk lehet egy olyan fordításból, amelynél sem az emberiség általános ismeretei, sem a szakismeret nincs teljesen felhasználva /amelyet tehát egy olyan gép adott ki, amelynek korlátozott általános emberi ismeretei vannak és gyenge "szakfelkészültsége"/.

A gyakorlatból is látszik /a "gyenge" fordítók gyakorlatából/, de elvileg is következik, hogy egy ilyen fordítást egy szakértő /ember/ az adott területen fel tud használni úgy, hogy nem ismervén a szöveg eredetének nyelvét, el tudja sajátítani azt az információt, amelyet az eredeti szöveg tartalmaz /természetesen ha a többi információ teljes egészében bekerült a "nyers" fordításba/.

A gépi fordítás és információ tárolás, valamint visszanyerés kapcsolata arra utal, hogy a gépi fordítás tényleges megoldása csak egyik része az általános információ tárolás és visszanyerés megoldásának. Ebből arra kell következtetnünk, hogy Magyarországon is teljes erővel be kell kapcsolódnunk az információ-tárolási és visszanyerés-

si kutatások körébe /annál is inkább, mert egyes kimagasló eredményeket ezen a téren éppen Magyarországon értek el, lásd pl. SZALAI Sándor munkáját./

.. ..

# BIBLIOGRÁFIA

- ÁBRAHÁM, S.: O principialno vozmozsnih perspektivah masinnovo perevoda, Computational Linguistic, 1964, 4 p. /nyomás alatt/
- ÁBRAHÁM S. - SALAPINA G.: On machine recognition of synonymy /nyomás alatt a "Proceedings of the 1962 Tihany Colloquium on the basis of mathematics mathematical machines and their application" c. könyvben/
- GARVIN, P.:/ed/ Natural language and the computer, Mc Graw-Hill, N.Y., 1963. 318.p.
- BAR - HILLEL, Y.: The Present Status of Automatic Translation of Languages, Advances in Computers, vol.1, F.L.Alt, ed. N.Y., Academic Press, 1960. 158-163.p.
- BAR - HILLEL, Y.: Four lectures on Algebraic Linguistics and Machine Translation, January 1963 /a revised version of a series of lectures given in July, 1962 before a NATO Advanced Summer Institute on Automatic Translation of Languages in Venice, Italy/
- OETTINGER, A.G.: The state of the art of automatic language translation: an appraisal. = Sprachkunde und Informationsverarbeitung, 2, 1963. 17-23.p.
- HOFFMANN, W./ed/ Digital information processors, Interscience Publ., N.Y., 1962
- Machine indexing: progress and problems, Washington D.C., 1961
- SZALAI, S.: Gépi kivonatkészítés. Elméleti és kísérleti adalékok a magyar nyelvű tudományos közlemények gépi kivonatolásának kérdéséhez. Bp. 1963. OMKDK, 69 p. /Időszertű műszaki dokumentációs kérdések 5./

A felvetett kérdések terén folytatott kutatások rendszeres ismertetését az olvasó a

Current Research and Development in Scientific Documentation  
N.S.P., Washington D.C.

című folyóiratban találja.



ABRAHÁM, S.: Some question of modern documentation

In this paper some questions of information processing and retrieval and the possibility of the use of computers in this field are discussed.

Information processing and retrieval must deal with: information classification, information heading, information extracting and information recognition.

For the information processing and retrieval by the aid of computers also an automatic pattern or/and speech recognition system is required.

In the information is in a foreign language, /machine/ translation is necessary. The goals and possibilities of MT are discussed, and only extract-translations are proposed as reasonable tasks. It is emphasised that MT is only a secondary question in machine information processing and retrieval, and that even the realization of MT does not add anything essential in the solution of information processing and retrieval. The need of developing vast research work in Hungary also in the field of theoretical and practical information processing and retrieval is also emphasised.

ooo

Абрахам С.:

Некоторые вопросы современной документации

В статье рассматриваются некоторые вопросы накопления и переработки информации и вопросы применения машин для решения этих вопросов. Задачами накопления информации считаются классификация информации, изготовление заголовков и кратких характерных содержаний, распознавание информации. Для машинного накопления и переработки информации нужна еще автоматическая вводная система /система автоматического распознавания звуковых и графических знаков/. В случае информации на иностранных языках встает еще вопрос /машинного/ перевода. Рассматриваются задачи и возможности МП и делается вывод, что имеет смысл только перевод заголовков и кратких характерных содержаний. Отмечается, что вопрос о машинном переводе является только частным вопросом теории накопления и переработки информации и даже ее решение не является основным в решении общего вопроса о машинном накоплении и переработке информации. В заключение отмечается необходимость развертывания и в Венгрии широких работ по теории и практике накопления и обработки информации.