Metainformációk előállítása: a kivonatolás szempontjai*

A gépi kivonatolással már azokban az „ősidőkben" elkezdtek foglalkozni a nyelvészek, programozók és számítástechnikai szakemberek, amikor az első nagyobb léptékű automatizált szövegfeldolgozások megkezdődtek. Az első figyelemre méltó tanulmány *H. P. Luhn* nevéhez fűződik, s azóta ez a terület mindig a figyelem középpontjában állt [1].

Több indítóoka is volt ezeknek a kutatásoknak; ezek közül csak kettő, amelyeket a nyelvészeti alapkutatások szükségessége (pl. morfológiai és szintaktikai elemzés) kapcsol szorosan össze:

- Már akkor is elrettentően hatott a folyamatosan keletkező és egyre növekvő információmennyiség, a bibliográfiai keresés egyre nehézkesebbé vált.
- Kezdetben nagy reményeket fűztek a gépi fordítás mielőbbi megvalósíthatóságához, ami a kezdeti sikerek után megakadt, hogy jóval később, szerényebb igényekkel és mértékben tovább folytassák.

A háttérben alapkutatások folytak több tudományágban [2], s a gyakorlati alkalmazások is fejlődtek:

1. A bibliográfusok és kivonatolók folyamatosan készítették a könyvek, cikkek, tanulmányok összefoglalóit, tökéletesítették a *KWIC* és *KWOC* indexeket, referatív kiadványokat jelentettek meg az információözön kezelésére, majd elkészültek a referatív kiadványok referatív kiadványai is. E területeken a vegyészek jártak az élen, más szaktudományok csak később zárkóztak fel.
2. A könyvtártudomány a gépi feldolgozás segítésére elkészítette a *MARC* szabványokat és alváltozatait (pl. *USMARC, FINMARC, UNIMARC, HUNMARC*). Ezek alapján választották ki azokat az adatleíró mezőket, amelyek összefoglaló neve *„Dublin Core Metadata Element Set"*. Ennek *DC 1.1* változatában 15 metaadatot megjelenítő, ismételhető azonosító mező szerepel, egyikük a *„Description"*, amely a forrás tartalmi összefoglalását tartalmazza. Ez a javaslat már az elektronikus dokumentumokra is kitekint [3].
3. A nyelvészek kvantitatív és kvalitatív nyelvészeti feldolgozásokkal bővítették ismereteinket a természetes nyelvek statisztikai szerkezetéről és annak működéséről a nyelvi leírás valamennyi szintjén. Az eredmények új szótárakban, szóelválasztási és helyesírás-ellenőrző programokban öltöttek testet. Ma már létezik korlátozott mértékű online fordítás néhány nagyobb idegen nyelv között, a szóelválasztás és helyesírás-ellenőrzés pedig minden valamirevaló szövegszerkesztő program szerves része.
4. A számítástechnikai szakemberek a hipertext alkalmazásával – kihasználva az egyre gyorsabb adatátviteli lehetőségeket – létrehozták a webet, s elkészültek azok a felhasználói programok is, amelyek nemcsak az itt található szöveges információk képernyőn való megjelenítését segítették, de felhasználóbarát eszközöket adtak arra is, hogy saját információinkat is könnyen hozzáférhetővé tegyük mások számára. Különösen fontos az adatbázis-kezelés fejlődésében a szövegekben való szabad keresés („ ") technikáinak kidolgozása, majd széles körű alkalmazása a webfelületen is. Idővel a figyelem a multimédiás megjelen(ít)és felé tolódott; ez jelentős mértékben a gyorsabb adatátvitelnek (is) köszönhető.
5. A matematikusok jól haladtak a gráfelméleti kutatásokban, amelyek eredményeit a Google 3 milliárd weblapot, 800 millió képet és naponta 200 millió kérést feldolgozó keresőmotorja nagymértékben hasznosítja. Jelentős eredmények születtek a szövegek matematikai-statisztikai leírásában, az automataelmélet a generatív grammatikában hasznosította eredményeit.
6. Az adatátviteli szakemberek ma már terabájtnyi információt tudnak átküldeni másodpercek alatt üvegszálas technikát alkalmazó szélessávú vonalakon az Atlanti-óceán alatt, vagy éppen mesterséges bolygók felhasználásával. További gyorsulás várható.

Ahogy fejlődtek a számítógépek, növekedtek a tárolási kapacitások és a processzorok sebessége, olyannyira miniatürizálódtak is: manapság az asztalunkon notebook formájában akkora számítástechnikai kapacitás lehet, mint az MTA 1980-ban beszerzett és minden célra használt IBM 3031 nagyszámítógép teljesítményének több ezerszerese.

A technikai fejlődéssel párhuzamosan új lehetőségek is keletkeztek: megjelent a világméretű pókháló, a *Web*, az *SGML* dokumentumleíró nyelv, és belőle megszületett a *HTML* protokoll. E tényezők együttesen lendületet adtak a szöveges feldolgozásoknak. Egyre nagyobb jelentősége lett a szö-

* *Maria Pinto* nyomán: Engineering the production of meta-information: the abstracting concern. = Journal of Information Science, 29. köt. 5. sz. 2003. p. 405–417.

vegekre, dokumentumokra vonatkozó metaadatoknak, s e célra bizonyos adatok automatikus előállításának.

Mivel tudjuk dokumentumainkat, szövegeinket jellemezni, a szöveges keresést segíteni? (Az alábbiakban nem térhetünk ki sok fontos részletre, kizárólag az eredeti cikk kérdéskörére szorítkozunk.)

1. *Semmivel.* Ez ugyan tipikus „matematikus" hozzáállásnak tűnhet, de okfejtésünkhöz szorosan hozzátartozik. Ez esetben fizikailag végig kell olvasni, vagy a számítógéppel olvastatni az egész szöveget, illetve az adatbázis minden mezőjét, s minden elemét össze kell hasonlítani a keresőkérdéssel. Mondanunk sem kell, hogy ez nem túlságosan hatékony és olcsó módszer.
2. *Szabad tárgyszavakkal.* Viszonylag kevés emberi munkaerő ráfordításával megoldható, hogy egy konkrét adatbázismezőt feltöltünk a dokumentumra jellemző szabadon választott tárgyszavakkal, pl. olyanokkal, amelyek az adott szövegben előfordulnak, s a keresőkérdést csak ennek tartalmával hasonlítjuk össze. A buktatók közül a szinonímiát, a hominímiát, az elkerülendő szakszavak használatát említhetjük, mindezek jelentősen lerontják találataink teljességét és pontosságát.
3. *Deszkriptorok használatával.* Itt már feltételezünk egy hátteret alkotó tezauruszt, amely tartalmazza a használható (és a tiltott) szavakat, s ezek között különböző viszonyokat is definiál (pl. rész-egész, alá- és fölérendeltség), s csak a megengedett deszkriptorokat szerepeltetjük a dokumentum leírásában, vagy az adatbázis egyik mezőjében. A keresőkérdést is e tezaurusz ellenőrzött terminusainak figyelembevételével fogalmazzuk meg, amivel jelentősen gyorsítható az összehasonlítás folyamata. A szabad tárgyszavakkal kapcsolatos buktatókat így elkerüljük ugyan, de jóval több emberi munkaerőt igényel a háttérben működtetett tezaurusz karbantartása.
4. *Tartalmi összefoglalóval.* Egyéb lehetséges nevei: annotáció, összefoglaló, tartalmi kivonat, ismertetés, áttekintés. A tartalmi összefoglalókat jó esetben olyan szakemberek készítik, akik értenek is valamennyire az adott tárgykörhöz. Ez a követelmény ma már egyre képtelenebb a tudomány és technika mai állása mellett. Itt lépnek színre és segítenek a számítógépek és a matematikai nyelvészeti módszerek. A keresés során a teljes dokumentum helyett csak a tartalmi összefoglalóval foglalkozunk, annak szavai, terminusai között keresünk valamilyen módon.

Ezzel megérkeztünk a szerző, *Maria Pinto* témájához, az automatikus kivonatolás problematikájához. A kérdéskör olyannyira fontos, hogy – egyebek mellett – a kutatók már a tartalmi összefoglalóknak a honlapokon való vizuális kiemelésével is foglalkoznak, ezzel segítve a felhasználókat a releváns információk azonosításában [4]. A halmozódó információmennyiség mérete szükségessé, a korszerű számítástechnika lehetővé teszi a kivonatolás jelentős mértékű automatizálását. Az eredmény minőségéről csak később szólok.

Maria Pinto nem említi ugyan a *Dublin Core* metaadatainak azonosító mezőit, de ő is alapvető tényezőnek tekinti a tartalmi összefoglaló (*abstract*) meglétét a dokumentum metaadatai között. A jelenleg működtetett számítástechnikai rendszerek lehetővé teszik az *információ* feldolgozását, de a pszichoszociológiai és pragmatikai háttér értékelését is feltételező *tudás*ét már nem. A számítógéppel ki tudjuk már választani a dokumentum jellegzetes mondatait, de jól használható tartalmi összefoglalót nem tudunk készíttetni vele.

Az elemzés során – Luhn idézett munkájától kezdődően – számos szellemes megoldást ismertet a szerző a gépi kivonatolás megvalósítására, a kapott eredmény javítására. A mai nyelvészeti kutatások kiterjednek a nyelvi jelenségek nagy tömegeinek statisztikai, kvantitatív jellemzése mellett a morfológiai és szintaktikai elemzésre, a szemantikai jelenségek tanulmányozására, valamint a beszédfordulatok, társalgási stratégiák pragmatikai feldolgozására.

Jó eredmények születtek a beszédfelismerés és szintézis terén is. A gépi fordításban is születtek új ötletek: ilyen pl. a kész mondatszerkezetek fordításainak tárolása, visszakeresése bonyolult fordítási algoritmus alkalmazása nélkül (l. *Prószéky Gábor* munkásságát).

Ha kilépünk egy-egy nagyon szűk szakmai szókincset használó szöveg keretei közül (pl. időjárásjelentések), jelentős problémák adódnak kisebb részben a szemantikai, nagyobb részben a pragmatikai szinten: szemantikailag a gép sok esetben fel tudja dolgozni a szöveg mondatait, de az ember számára „természetes" háttérismereteket, a mögöttes szándékokat, a kommunikációs partnerek közötti elvárásokat már nem tudja értelmezni és visszaadni.

Pinto megkülönbözteti a tartalmi kivonat három fajtáját és létrehozási módszereiket:

- *Extract* vagy szövegkivonat. Statisztikai módszerekkel kiválasztja a kulcsszavakat, címszavakat, a szöveg kiemelt helyein vagy mondataiban található releváns szavakat, s ezeket használja fel, esetleg egy kiegészítő szótár (*Word Control List*) alapján súlyozva fontosságukat. Ezek között az egyik legjobb módszer a szóláncok szövegbeli azonosításán és használatán alapul, míg a másik meghatározott gyakorisági küszöbök közötti szavakat és szókapcsolatokat használ (a leggyakoribbak és a csak egyszer előforduló szavak a szöveg jellemzésére nem megfelelőek a keresés segítésének szempontjából).
- *Summary* vagy összefoglaló. A dokumentum szövegének fizikai megjelenése, vagyis az ún. felszíni szerkezete adja az alapot, ezt egészítik ki retorikai és kognitív struktúrák (minták, angolul: *frame*) felismertetésével, ezzel is javítva az összefoglaló nyelvi minőségét. Az *AltaVista* a keresés segítésére gyakran használ természetes nyelven megadott mintákat, a felhasználónak csak ki kell választania a neki megfelelőt. Az *Ask Jeeves* az emberi keresések gyakorlatát és eredményét tárolja adatbázisában. Ezután szakemberek tekintik át a webforrásokat, s tudásbázist építve mintákban tárolják azokat a forrásokat, amelyek az általános kérdések megválaszolására használhatók. A kognitív struktúrákat használja több szűk körben régóta használatos rendszer, pl. SUMMONS, FRUMP, SUSY, SCISOR.
- *Abstract* vagy tartalmi kivonat. A szerző szerint ez a legmagasabb szintű eredményt adja, de a számítógép által valamilyen módon előkészített szöveget kizárólag emberi közreműködéssel lehet koherenssé, emberi fogyasztásra is alkalmassá tenni.

Általánosságban elmondható, hogy a sok automatizált módszer közül a legkifinomultabb sem kielégítő az időráfordítást, feldolgozási költségeket és – ami számunkra talán a legfőbb – a minőséget tekintve. A tartalmi kivonat előállítása olyan kognitív és pragmatikai tevékenység, amelyet nagymértékben befolyásol a kontextus, s magának a tartalmi kivonatnak a használata is kontextusfüggő. Figyelembe kell venni a felhasználót, feltételezhető szándékait, a dokumentumforrás típusát, a tartalmi kivonat célját, valamint egyéb „külső", magából a dokumentumból nem következő körülményeket. Ezért ugyanabból a dokumentumból sok-sok tartalmi kivonat készíthető.

A tartalmi kivonatok (abstract) készítésének teljes automatizálása irreális álom, de szövegkivonatok (*extract*) vagy összefoglalók (*summary*) készítése automatizálható ugyan, azonban használatra nem megfelelő minőségben. Ezért a szerző azt az elképzelést tartja elfogadhatónak, hogy a tartalmi kivonat elkészítésének automatizált folyamatában emberi közreműködést is alkalmazzunk (hasonlóan a számítógéppel támogatott „gépi fordításhoz"). Ezt a működési módszert valósítja meg a TEXNET kísérleti rendszer, amely tartalmi kivonatkészítést segítő munkaállomásként képzelhető el: bizonyos feladatokat az ember végez, az összes többi feladatot pedig a számítógép a háttérben.

Elméleti szinten az értékek átértékelése folyik, amit a mindennapi gyakorlat ösztönöz. Egyre kevésbé fontosak az információ-visszakeresés logikai-szemantikai alapjai, az indexelési módszerek, az információs keresőnyelvek problémái. A működő rendszerek technológiai kérdései viszont sokkal fontosabbak, több figyelem jut az új algoritmusok kifejlesztésére, a programokra, legalábbis ez látható a keresőmotorok dinamikus fejlődéséből. Voltak olyan kutatások, amelyek arra jutottak, hogy a tezauruszos keresés a szabad szövegessel szemben nem ad lényegesen pontosabb eredményt, vagyis a teljesség igénye csak a pontosság rovására tud teljesülni és fordítva. Következésképpen nem is érdemes fejleszteni a kifinomult technikákat, hiszen a gyors szabad szöveges kereséssel működő rendszerek a jelenlegi számítógépeket figyelembe véve olcsók, jól kezelhetők, mégpedig távolról, a felhasználók saját számítógépéről.

Ezen a ponton nem hanyagolható el, hogy a Pinto cikkében részletezett kutatások döntő többsége az angol nyelvre vonatkozik, amely nyelvtipológiailag *analitikus* nyelv. Nem teljesen szakszerűen fogalmazva ugyanannak a szótőnek (mondjuk „szótári szó"-nak) az angol nyelvben nagyon kevés szóalakváltozata van a konkrét szövegekben (azaz gyakorlatilag nincs ragozás). A dokumentumok és természetes nyelvi szövegek feldolgozásával német és orosz nyelvterületen is sokat foglalkoztak, de ezekben a főként *flektáló* típusúnak tartott nyelvekben sem olyan nagy mennyiségű az azonos szótövekhez tartozó különböző szóalakok száma, mint az ún. *agglutináló* nyelvekben, mint a magyar, mongol vagy török. Mindez azt jelenti, hogy bármilyen, a magyar nyelvre vonatkozó szövegfeldolgozásban kiemelt szerepe van a *morfológiai elemzés*nek, hiszen egy-egy szótári szavunknak főnevek esetében közel ezer különböző alakja kerülhet

elő bármikor a konkrét feldolgozott szövegekben (bár az is igaz, hogy ezek a szóalakok messze nem azonos valószínűséggel fordulnak elő). A Microsoft Word magyar helyesírás-ellenőrzőjének néha elég furcsa javaslataiból is tapasztalhatjuk, hogy nem könnyű még ez a nyelvleírás szempontjából viszonylag egyszerű feladat sem. A *szintaktikai elemzés*ben is nehézséget jelent a magyar nyelv szabad szórendje az igencsak kötött angol szórendhez képest. Ezért bizonyos fentebb felsorolt technikák nehezen, vagy nem alkalmazhatók az agglutinatív nyelvekre.

Vannak olyan magyar vonatkozású háttérmunkálatok, amelyekre keveset és kevesen (vagy éppen egyáltalán nem) hivatkoznak. Ilyenek a *Papp Ferenc* által szerkesztett *Szóvégmutató szótár* (Bp., Akadémiai Kiadó, 1968), a *Füredi M.–Kelemen J.: A mai magyar szépprózа gyakorisági szótára* (Bp., Akadémiai Kiadó, 1989), s e kettőnek számítógépes adatbázisai (Debreceni tezaurusz, GyakNagy), amelyeknek jó időben *„public domain"*-nek nyilvánított, több száz emberévnyi munkával elért eredményeit messzemenően bedolgozták termékeikbe a magyar helyesírás-ellenőrző programok készítői. Ezek az 1960-as, 1970-es években manuális kódolással elvégzett alapkutatások és számítógépes feldolgozásaik adtak megbízható hátteret a mai magyar nyelv szótári szerkezetének, és a nagyobb lélegzetű irodalmi szövegek statisztikai, morfológiai, esetenként szintaktikai és szemantikai sajátosságainak megértéséhez [5].

Bár a nyolcvanas évek elején teljesen megszűnt a kvantitatív nyelvészeti kutatások támogatása, ma már remény van arra, hogy a Nyelvtudományi Intézetben összegyűlt beszélt nyelvi anyagok, valamint az Akadémiai Nagyszótár céljaira kódolt anyagok alapján – felhasználva az azóta is fejlődő nyelvészeti elemző technikákat – különböző műfajokra, közöttük a tudományos és tudománynépszerűsítő szövegekre is lehessen következtetéseket levonni, ezzel is segítve a könyvtári dokumentum-visszakereső szolgáltatásokat.

Irodalom

[1] LUHN, H. P.: A statistical approach to mechanised encoding and searching of literature information. = IBM Journal Research and Development, 1. köt. 4. sz. 1957.
LUHN, H. P.: The automatic creation of literature abstracts. = IBM Journal Research and Development, 2. köt. 2. sz. 1958. p. 159–165.
STEVENS, M. E.: Automatic indexing – A state of the art report. National Bureau of Standards, Washington, D. C., 1965.
FANGMEYER, H.: Semi-automatic indexing: state of the art. Paris, 1974.

[2] GUIRAUD, P.: Les caractères statistiques du vocabulaire. Essai de méthodologie. Presses Universitaires, Paris, 1954.
HERDAN, G.: Type/token mathematics. 's Gravenhage, 1960.
HERDAN, G.: Quantitative linguistics. London, 1964.
NAMBIAR, K. K.: Theory of search engines. 2001. január 8.
SHANNON, C. E.: The mathematical theory of communication. University of Illinois Press, Chicago, 1963.

[3] Dublin Core metadata element set reference description, version 1.1. 1999. 07. 02. *http://purl.org/dc//documents/proposed_recommendations/pr-dces-19990702.htm*

[4] HUSSAM, A. et al.: Semantic highlighting: enhancing search engine display and Web document interactivity. = Human-Computer Interaction – INTERACT – 1999. II. köt. Szerk.: S. Brewster–A. Cawsey–G. Cockton. The British Computer Society. IFIP TC. 1.3. 1999. p. 85–86.

[5] FÜREDI M.: Összesített adatok a szépprózai gyakorisági szótárról. Igék és igeszármazékok. = Magyar Nyelv, 82. köt. 2. sz. 1986. p. 190–198.

(Füredi Mihály)

Nyomtatott és elektronikus folyóiratok kezelési költségeinek összehasonlító vizsgálata

Sok felsőoktatási és tudományos könyvtár van azon az úton, hogy folyóirat-gyűjteményét nyomtatott formátumról elektronikusra váltsa. Ennek következtében – kivéve a legnagyobb könyvtárakat – az elektronikus folyóirat-gyűjtemények mérete messze meghaladja a nyomtatottaknál valaha is tapasztalt nagyságot. Ez természetszerűen kihat a folyóiratokkal végzett könyvtári műveletekre, és a személyzettel szemben támasztott követelményekre is. Rendszerszerű megközelítésben különös