

Programcsomag információkinyerési kutatások támogatására

Alexin Zoltán¹, Gyimóthy Tibor¹

Szegedi Tudományegyetem, TTK, Informatikai Tanszékcsoport,
Szeged Árpád tér 2., e-mail:{alexin,gyimothy}@inf.u-szeged.hu

Kivonat A publikációban bemutatásra kerül egy információkinyerési kutatásokat támogató programcsomag, amelynek moduljai a nyers szöveg beolvasásától kezdve a végeredmény webes megjelenítéséig minden szükséges funkciót megvalósítanak.¹

Kulcsszavak: információkinyerés, természetesnyelv-feldolgozás, felszíni szintaktikai elemzés

1. Bevezetés

Az információkinyerés (IE, Information Extraction) technológiájának kutatása dinamikusan fejlődő terület a természetesnyelv-feldolgozásban. Az Interneten megjelenő hatalmas információtömeg gépi feldolgozása és a kívánt információ tömör formában történő összegyűjtése napi szükséglet, amelyre a gazdaság, a tudomány, a politika, de akár a hírszerzés területén is van igény. Míg az információ visszakeresés (IR, Information Retrieval), amely a webes kereső programok jellemző tevékenysége, arra irányul, hogy a felhasználó igényeinek megfelelő dokumentumokat változatlan formában bocsássa rendelkezésre, addig az információkinyerés célja a megtalált dokumentumokban a lényeges információ megjelölése, majd összegyűjtése. A számítógéppel támogatott szövegtömörítés, kivonatolás és az információkinyerés szoros kapcsolatban áll egymással.

2. Matematikai képletek és formulák

Theorem 1. *Tegyük fel, hogy $H = C^2$ és (a_∞, b_∞) -subquadratkus a végtelenben. Legyen ξ_1, \dots, ξ_N egyensúlyi pontok, azaz a $H'(\xi) = 0$ megoldásai. Jelölje ω_k a $H''(\xi_k)$ legkisebb sajátértékét, és legyen:*

$$\omega := \text{Min } \{\omega_1, \dots, \omega_k\} . \quad (1)$$

Ha:

$$\frac{T}{2\pi} b_\infty < -E \left[-\frac{T}{2\pi} a_\infty \right] < \frac{T}{2\pi} \omega \quad (2)$$

¹ A szerzők köszönetüket fejezik ki az Oktatási Minisztériumnak, amely az NKFP 2/017/2001 projekt keretében az itt ismertetésre kerülő kutatást támogatta.

akkor ψ minimalizálása egy nem-konstant T -periodikus megoldást eredményez \bar{x} .

Emlékeztetünk rá még egyszer, hogy $\alpha \in \mathbb{R}$ esetén az $E[\alpha]$ egészrészén az $a \in \mathbb{Z}$ értjük, ahol $a < \alpha \leq a + 1$. Például, ha $a_\infty = 0$, akkor a tétel azt állítja számunkra, hogy \bar{x} létezik és nem konstans, feltéve, hogy teljesül:

$$\frac{T}{2\pi} b_\infty < 1 < \frac{T}{2\pi} \quad (3)$$

vagy

$$T \in \left(\frac{2\pi}{\omega}, \frac{2\pi}{b_\infty} \right) . \quad (4)$$

2.1. Az információkinyeréshez kapcsolódó modulok fejlesztését támogató adatbázis: a Szeged Korpusz 2.0

Az Oktatási Minisztérium által támogatott IKTA 27/2000 projekt keretében készült el a Szeged Korpusz 1.0-s változata[1], amely egy szófajlag elemzett, majd kézzel egyértelműsített adatbázis volt. Ezt az információkinyerési kutatások támogatására az MTA Nyelvtudományi Intézettel és a MorphoLogic Kft.-vel közös konzorcium jelentősen továbbfejlesztette. A Szeged Korpusz újabb változata ² hat különböző témakörben gyűjtött, összesen 1,2 millió szót tartalmazó, számítógéppel feldolgozható szöveg. Ennek az állománynak egy mintegy 200 ezer szavas részét képezi a bevezetőben már említett 6453 MTI rövidhírt tartalmazó anyag.

```
<xml>
  <sentence id="1.1">
    <word>A<mscat>NE</mscat></word>
    <word>kutya<mscat>FN</mscat></word>
    <word>ugat<mscat>IGE</mscat></word>
    <punctuation>.</punctuation>
  </sentence>
</xml>
```

1. ábra. Egy XML állomány részlete

2.2. A beolvasott szöveg szegmentálását végző modul

A beolvasott szöveg XML adatbázissá alakítása és az alapvető metainformációk (fejezet-, bekezdés-, mondat-, szóstruktúra) meghatározása a feldolgozás első lépése[2] [5]. A természetes nyelvi szövegekben számos különböző fajta szó

² SZTE, Informatikai Tanszékcsoport, Nyelvtechnológiai Csoport: <http://www.inf.u-szeged.hu/hlt>

jellegű, de a szótárakban nem szereplő lexikai elem található (szám, dátum, gépkocsirendszám, e-mail cím, stb.), amelyek felismerésére valamint a mondat-határok megállapítására egy formális-nyelvi eszközöket alkalmazó modul készült. A modul a GNU Flex ³ reguláris automatagenerátor eszközt használja. Ebben reguláris kifejezések írják le az ún. tokeneket (2. ábra). A *flex* program a reguláris kifejezésekből C programot készít, amely a szegmentáló modul magját alkotja. A mondatokra és a szavakra bontás hatásfoka igen jó, a hibásan felismert tokenek aránya nem több, mint 0,5%.

```
/* ponttal tagolt számok, pl. 12.000 */
NUMDOT [0-9]{1,3}("."[0-9]{3})+
NUMDIGIT ([0-9]+",")?[0-9]+
```

2. ábra. Reguláris kifejezések a Flex definíciós fájljában

3. A felhasználói felület és a webes megjelenítő modul

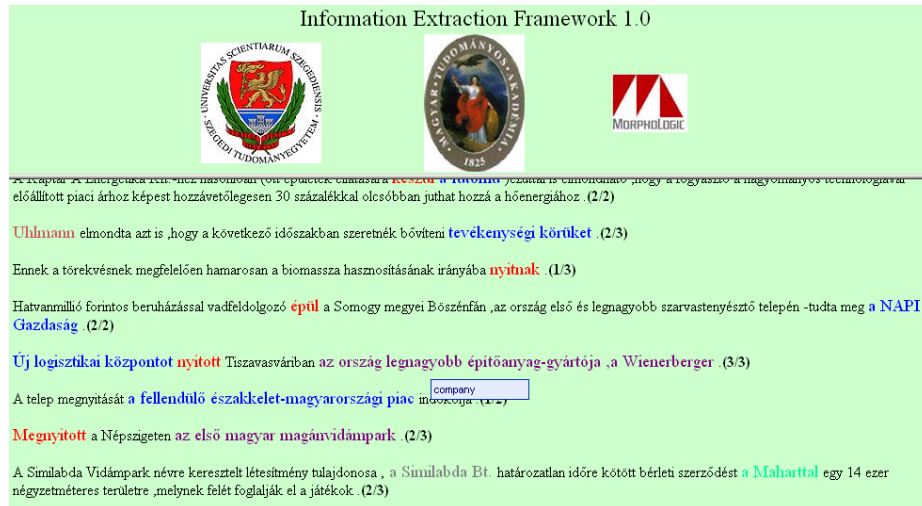
A programcsomagban az utolsó modul a felhasználói felület, amely egy HTML nyelvű weblapot készít. Ez egyrészt tartalmazza a beolvasott nyers szöveget, másrészt a programcsomag által hozzáadott metainformációkat. Ez utóbbiakat grafikus eszközökkel jeleníti meg. A mondatokban azonosított szereplők különböző színekkel, a szerepek nevei a weblapon lebegő üzenetablakokban (*tooltipekben*) jelennek meg. A mondatokra illesztett szemantikus keret összes szereplőjének száma és az azonosított szereplők száma a mondatok után található. A 3. ábrán a webes megjelenítő modul egy képernyője látható.

1. táblázat. Ezt a példát a *The T_EXbook*, 246. oldaláról vettük

Év	A világ népessége
8000 B.C.	5,000,000
50 A.D.	200,000,000
1650 A.D.	500,000,000
1945 A.D.	2,300,000,000
1980 A.D.	4,400,000,000

Egy alternatív megjelenítő modul az eredményeket nem weblapon, hanem Excel-ablakban jeleníti meg. Az azonos eseményeket egy munkalapon, a mondatokat egy-egy sorban, az azonos szereplőket pedig azonos oszlopokban jeleníti meg. Ez a táblázat további feldolgozások kiindulópontja lehet.

³ A GNU Flex honlapja: <http://www.gnu.org/software/flex>



3. ábra. A webes megjelenítő modul egy képernyője

4. Eredmények

A cikkben bemutatott programcsomag tesztelésére a kutatók egy keretrend- szert (benchmark) készítettek. Ez kézzel előre annotált, a rendszer számára ismeretlen mondatokat tartalmaz két előre kiválasztott témakörben: a tulajdonos-váltás és az új telephely nyitása témakörében.⁴ Erre a két témakörre megfelelő számú szemantikus keret-definíció és rövidhír állt rendelkezésre. A teszt-mondatok szöveges alakjára lefuttatták a programcsomag egyes komponenseit, majd összehasonlították a kézzel készített és a gép által előállított két állomány metainformációit. Tekintve, hogy a programcsomag több elemből áll, az egyes modulok hibája kumulálódik a végeredményben. Az eredmények megbízhatóbb értékelése érdekében arra is van lehetőség, hogy az egyes modulok által adott részeredményeket külön értékeljék, és összehasonlítsák az etalonfájllal.

5. Köszönetnyilvánítás

A szerzők ezúton fejezik ki köszönetüket az OM NKFP 2/017/2001 projektbeli konzorciumi partnereiknek, az MTA Nyelvtudományi Intézet Korpusznyelvészeti Osztályának és a MorphoLogic Kft.-nek, akikkel a tudományos és szakmai kapcsolatokon túl szoros, személyes kapcsolatot alakítottak ki.

Hivatkozások

1. Alexin Z., Csirik, J., Gyimóthy, T., Bibok K., Hatvani, Cs., Prószték, G., Tihamy, L.: Manually Annotated Hungarian Corpus. in Proc. of the Research Note

⁴ A benchmarkban 176 rövidhír, illetve 285 mondat szerepel.

- Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics EACL'03, Budapest, Hungary, 53–56 (2003).
2. Bibok K.: A szóról és a szófajokról (a számítógépes nyelvfeldolgozás kap- csán), Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003), Szeged, Magyar- ország, 31–36, (2003).
 3. Farkas R., Konczer K., Szarvas Gy.: Szemantikus keretillesztés és az IE rendszer automatikus kiértékelése Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2004), beküldve, Szeged, Magyarország, (2004).
 4. Hócz, A.: Noun Phrase Recognition with Tree Patterns elfogadva az Acta Cyber- netica c. lapban történő megjelenésre (2004).
 5. Mihácz András, Németh László, Rácz Miklós: Magyar szövegek természetes nyelvi előfeldolgozása Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003), Sze- ged, Magyarország, 38–43, (2003).
 6. Prószéky G.: Automatikus információszerzés gazdasági rövidhírekből. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003), Szeged, Magyarország, 161– 166, (2003).
 7. Prószéky G.: Automatikus információszerzés gazdasági-politikai rövidhírekből. VIII. Országos (Centenárium) Neumann Kongresszus kiadványa, Budapest, Ma- gyarország, 359–367, (2003).