Bednarik László
Eszterházy Károly Főiskola, Comenius Kar, Reál Tudományok Intézete
bednarik@ekfck.hu

SZEMI-AUTOMATIZÁLT KÉRDÉSGENERÁLÓ RENDSZER FUNKCIONÁLIS MODULJAI

1. Bevezetés

A mesterséges intelligencia az 1950-es évek közepén jelenik meg, majd a különböző adatbázisok (integrált adatraktár, hierarchikus, hálós és relációs adatmodell, SQL) fejlődésén keresztül eljutunk az adatbányászatig. Az adatbányászat feladata elsősorban az adatbázisokban, adattárházakban lévő strukturált adatok feltárása. A 1990-es évek végén az adatbányászatból fejlődik ki a szövegbányászat. A szövegbányászat célja a szöveges dokumentumok elemzésére és feldolgozására szolgáló algoritmusok kifejlesztése, amelyek az emberi nyelv tudását összekapcsolja a számítógép nagy feldolgozási kapacitásával[1].

Napjainkban a hagyományos oktatási keretrendszerek mellett megjelennek az adat- és tudásbázisok, szemantikai adatbázisok. Ezen adatbázisok tudásanyagára építve rugalmasan kezelhető a tananyag.

A kidolgozott szemi-automatizált kérdésgeneráló rendszer (AQG) képes annotált magyar nyelvű szöveges dokumentumból, választható mondattípusok alapján, feleletválasztós és kiegészítős kérdések automatikus előállítására.

2. Háttér

Az automatizált kérdésgeneráláshoz kapcsolódó vizsgálatok az 1990-es évekig nyúlnak vissza. A korábbi kutatások a kérdésgenerálás szemantikai aspektusára fókuszáltak, amelyek módszertani alapként szolgálnak a mai automatizált rendszerekhez.

Az AQG létrehozásának kutatásában a domináns megközelítést Miller[2] javasolta, melyben a WordNet lexikális tudásbázisra alapozva hat kérdéstípus különböztethető meg: definíció, szinonim, antonim, hipernim, hiponim és feleletválasztós. Sumita 2004-ben javasolt egy automatikus generáló módszert a feleletválasztós kérdések előállítására[3]. A mondatok kiválasztása, az üres helyek és a hiányos mondatok meghatározása a gépi tanulási módszerek felhasználásával valósult meg. Nielsen és szerzőtársai[4] megalkottak olyan modellt, amely a kérdésgeneráláshoz szükséges adatokat és műveleteket felhasználták a rendszer elkészítéséhez. Gütl és munkatársai 2008-ban elkészítettek egy modulból álló automatizált kérdésgeneráló rendszert. A kifejlesztett prototípus eredményein alapulva 2010-ben továbbfejlesztették a rendszert, amely már

[1] Fajszi et al., 2004
[2] Miller, 1995
[3] Sumita et al., 2004
[4] Nielsen et al., 2008

három modulból állt: előfeldolgozási, fogalmak kinyerésére szolgáló modul és a kérdésgeneráló modul[5].

3. Az AQG modellje

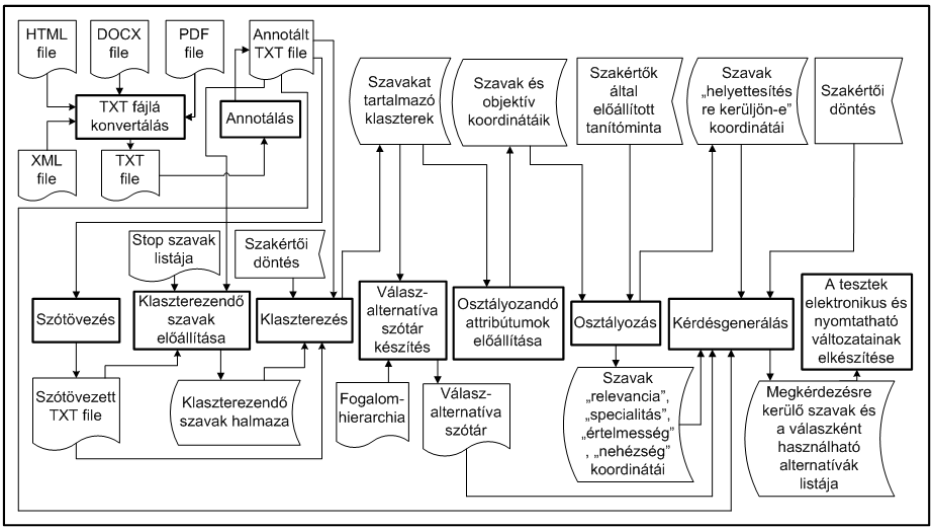
A szakirodalomban meglévő automatizált kérdésgeneráló rendszerek jelentős része idegen nyelvű szövegek kezelésére fejlesztették ki. A magyar nyelvű szöveges dokumentumok kísérleti stádiuma miatt egy saját fejlesztésű, a gyakorlati életben is alkalmazható rendszert dolgoztam ki. Az automatizált kérdésgeneráló rendszerrel szemben támasztott főbb követelmények: magyar nyelvű támogatás, ingyenes nyelvtani elemző használata, automatizált gépi tanulás, tesztkérdések készítése annotált szövegből, válaszlehetőségek előállítása belső, saját fejlesztésű szemantikai adatbázisból, feleletválasztós és kiegészítő feladatok támogatása, elektronikus és nyomtatható feladatlapok készítésének.

A megvalósított automatizált kérdésgeneráló rendszer több alrendszerből épül fel. Minden modul bemeneti, illetve kimeneti interfészt tartalmaz. A bemeneti interfészek meghatározzák azokat az adatokat, amelyeket a modulok igényelnek a feladatuk elvégzéséhez, a kimeneti interfészek definiálják azokat az adatokat, amelyeket a moduloknak szolgáltatniuk kell. A rendszerterv elkészítése során szabványos modellelemeket használtam: szöveges, illetve bináris fájlban tárolt adat, belső adatreprezentációban tárolt adat, jól definiált feladatot ellátó és döntési folyamatot reprezentáló modul.

Az automatizált kérdésgenerálás feladatát ellátó rendszermodell funkcionális rendszertervét az *1. ábra* szemlélteti[6]. A cikk alfejezetei adatáramlás alapján mutatják be a funkcionális rendszerterv-modell működését.

[5] Gütl et al., 2011

[6] Bednarik, 2012



1. ábra: A rendszermodell funkcionális rendszerterve

3.1. Előfeldolgozó modul

A kérdésgeneráló rendszer első lépése az előfeldolgozás, melynek célja a dokumentumokat olyan alakra hozni, melyben a klaszterezési és osztályozási feladatok hatékonyan elvégezhetők. Az annotációs modul célja, hogy az egyes mondatokhoz hozzárendeljen egy-egy szerepkört. Az annotáció megvalósítása a kérdésgeneráló rendszer első fázisában manuálisan történt, a fejlesztés jelenleg az automatikus szerepkijelölés felé halad. Az előfeldolgozást végző modul támogatja a bemenetként szolgáló nagyméretű szöveges dokumentum mondatainak annotációval megvalósított szűrését, illetve a mondatok kategóriákba sorolását (fogalom, definíció, kijelentő mondat). A megvalósított modell előfeldolgozó modulja alkalmas DOC, DOCX, RTF, HTML és PDF formátumban kódolt dokumentumoknak, bemeneti adatként való kezelésére.

3.2. Szótövező modul

A szótövezés a szavak szótőre redukálását jelenti. A szótövezés révén jelentősen redukálható a kezelt, felismerendő szavak halmaza, hiszen a magyar nyelvben egy alapszónak 20–50 ragozott alakja is megjelenhet a szövegben. A kereskedelemben több szótövező alkalmazás megtalálható, ezek közül csak néhány ingyenes (szószablya, magyarlánc).

A keretrendszerben egy ingyenes nyelvi elemző, a Szószablya keretrendszer[7] került beépítésre, mert ez biztosította a legpontosabb elemzési lehetőségeket a különböző szöveges dokumentumok esetén. Ennek algoritmusa a Porter algoritmus[8] adaptálásával működik. A módszer fő előnye a nagyfokú gyorsaság. Az elemzett szavakról a

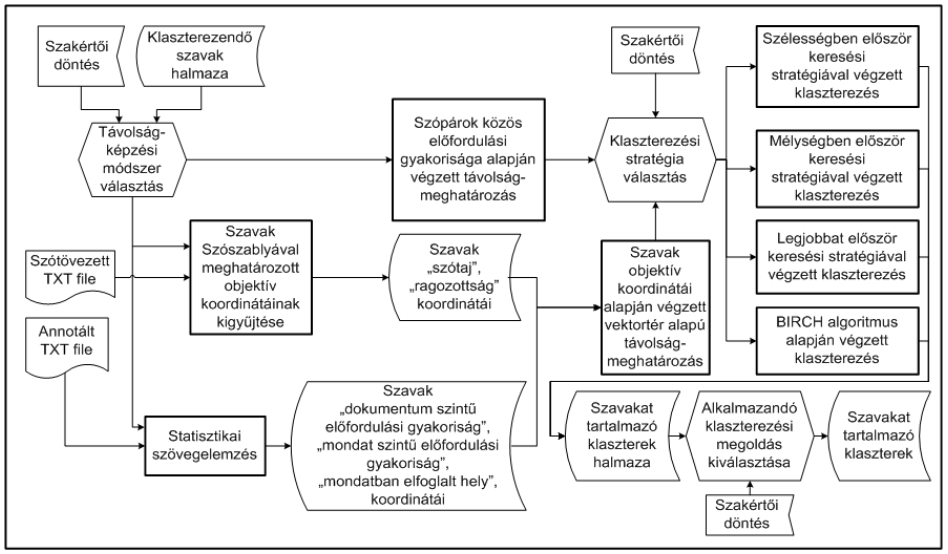
[7] Németh, 2003
[8] Tikk, 2007

következő jellemzők kerülnek előállításra pontosvesszőkkel szeparált szöveges fájlban: szóalak, szótő, gyakoriság, szótagszám, elemzés, szófaj.

3.3. Klaszterező modul

A klaszterezés célja a dokumentum szavainak csoportokba szervezése, ahol az azonos csoportba sorolt szavak egymáshoz minél hasonlóbbak, míg a különböző csoportba soroltak egymástól különbözőbbek legyenek[9].

A klaszterezés alapvető fontosságú feladat az automatizált kérdésgenerálás során. A klaszterezést végző modul működéséhez, bemeneti információként kapja a klaszterezendő szavak halmazát. Ezt a halmazt az előző modulok állítják elő, a kérdésgenerálásra alkalmazott dokumentumból kiszűrve az annotációval nem rendelkező mondatokat, valamint a stopszavak listáján szereplő szavakat. A megmaradt szavak rendezetlen halmaza adja a klaszterezést végző modul egyik bemenetét. Ezután szakértői döntés alapján kerül kiválasztásra az alkalmazandó szótávolság-képzési módszer. A klaszterezést végző modul rendszertervét a *2. ábra* szemlélteti.



2. ábra: Klaszterező modul

A dokumentumok szavainak klaszterezésére a szógyakoriság-alapú távolságképzést fejlesztettem ki. A szavak közös előfordulási gyakoriságára épülő távolság-meghatározási koncepció értelmében két szó távolsága azon mondatok számával definiálható, melyekben mindkét szó egyszerre szerepel. Az így meghatározott távolságadatok távolságmátrixszal írhatók le. A szavak távolságadatainak meghatározása két kifejlesztett módszer alapján valósult meg. Az egyik módszer a „Távolságmátrix újraszámolása minden klaszterösszevonást követően" történik, ennek implementálása a

[9] Bodon, 2010

mélységi keresési stratégiával valósult meg. A másik módszer, amikor az algoritmus a „Kiinduló távolságmátrixot használja a teljes klaszterezési folyamat” során. Ennek a megvalósítása a mélységkorlátos mélységben először keresési stratégia alkalmazása.

A távolságmátrixnak a sorai és oszlopai a dokumentum szavaival kerülnek indexelésre. A mátrix i sorában és j oszlopában szereplő érték a (1) összefüggés alapján került meghatározásra:

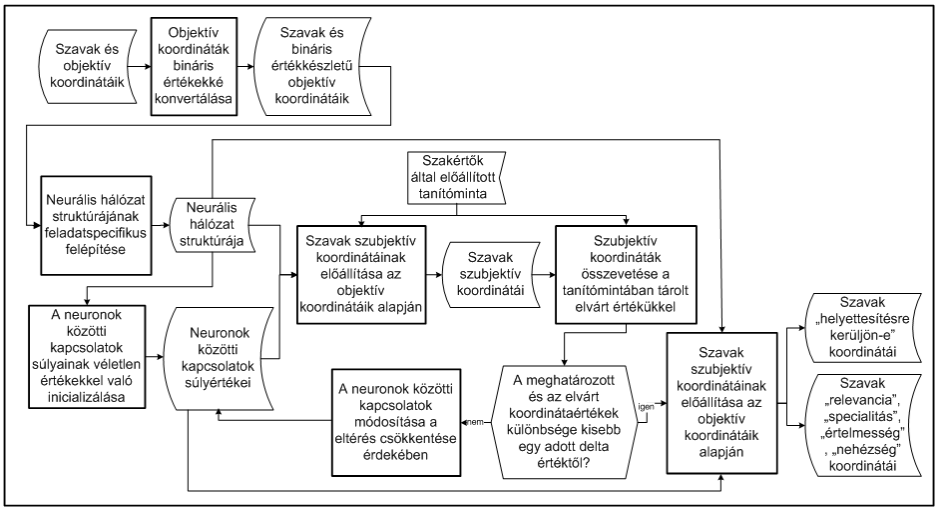
$$S_{i,j} = f_{i,j} \,/\, \max(f_i, f_j) \qquad (1)$$

ahol: S_{ij} az i és j szó távolságviszonyát reprezentáló [0, 1] intervallumba eső számérték; f_{ij} a dokumentum azon mondatainak száma, melyekben az i és a j szó is szerepel; f_i a dokumentum azon mondatainak a száma, melyekben az i szó szerepel; f_j a dokumentum azon mondatainak a száma, melyekben a j szó szerepel.

A kiinduló távolságmátrixra épülő klaszterezés feladatspecifikus továbbfejlesztése az „Egységes átmérőjű koronggal végzett klaszterezés”. Az elkészített algoritmus ennek a módszernek a QTC (Quality Treshold Clustering, minőségi küszöbérték klaszterezés) algoritmus adaptálásával valósítottam meg, melyben a klasztereket előre definiált egyenlő sugarú körökkel modelleztem. A klaszterezést végző optimalizációs algoritmus célfüggvényét a klaszterszám minimalizálásával, a korlátfeltételét pedig a dokumentum minden szavának legalább egy klaszterhez tartozásával definiáltam. Az algoritmust, a legjobbat először keresési stratégia alkalmazásával valósítottam meg.

3.4. Osztályozó modul

A modul feladata a dokumentum szavainak objektíven mérhető koordinátái alapján a szavakhoz definiált szubjektív koordináták meghatározása. A szavaknak a nyelvészeti, valamint statisztikai módszerekkel objektíven mérhető tulajdonságait a szavak objektív koordinátáinak nevezzük. A szubjektív koordináták a szavaknak azokat jellemzőit jelölik, melyek az embernek az adott szóra vonatkozó megítélését fejezik ki. A kidolgozott mintarendszerben a következő szubjektív koordináták kerültek bevezetésre: nehézség, relevancia, specialitás, értelmesség. Az osztályozást végző modul rendszertervét a *3. ábra* szemlélteti.



3. ábra: Osztályozó modul

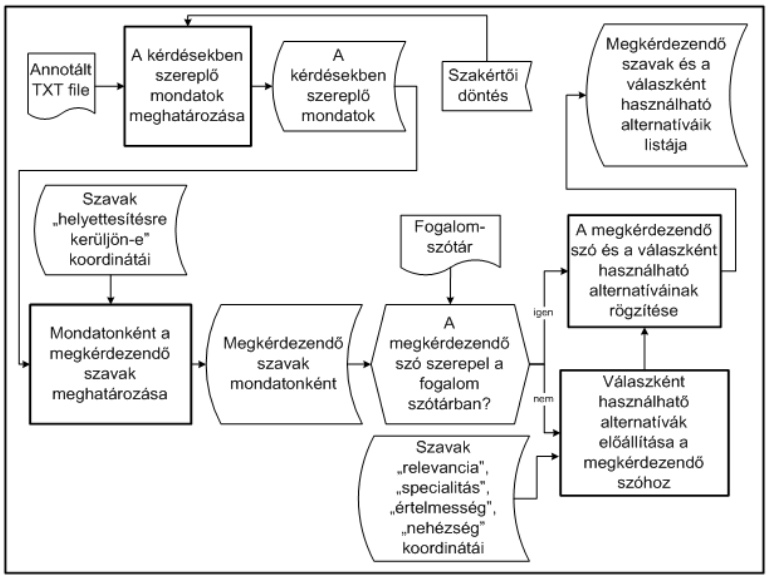
Az osztályozási feladatot neurális hálózat alkalmazásával végeztem el. Az osztályozást végző modul bemeneti információként kapja az előző modulok által előállított szavakat és a hozzájuk tartozó objektív koordinátákat. A szavak objektív koordinátái részben megegyeznek a klaszterezést végző modulban előállított objektív koordinátákkal, de az osztályozás megkezdése előtt ezek a koordináták kiegészülnek a klaszterezés eredményének ismeretével nyerhető újabb információkkal. Az új objektív dimenziók: a szót tartalmazó klaszter sorszáma, valamint a szó átlagos távolsága a szót tartalmazó mondat többi szavától. Ezekkel együtt az osztályozás bemenetén minden szó egy hétdimenziós objektív térben kerül elhelyezésre.

Az osztályozási feladat magját egy háromrétegű előrecsatolt neurális hálózattal valósítottam meg. A hálózat betanítását a felügyelt tanítási módszer alapján végeztem el. A neurális háló bemeneti rétegében lévő neuronok számának pontos meghatározására külön algoritmust fejlesztettem ki. Az algoritmus feltárja az objektív koordináták által felvehető értékek halmazát és minden objektív koordináta minden lehetséges értékéhez külön neuront vesz fel a bemeneti rétegben. A neurális hálózatok építésének egyik kulcskérdését jelenti a belső rétegben lévő neuronok számának meghatározása. Az optimális értéktől alacsonyabb számú belső neuron alkalmazása esetén a hálózat nem lesz képes a feladat megtanulásához szükséges mennyiségű információ tárolására, illetve a neurális háló betanultsági szintje nem képes elérni az elvárt értéket. Az optimális értéktől több belső neuron használata esetén a neurális hálózat általánosító képessége csökken, így a hálózat csökkent mértékben képes a tanításra használt mintákban rejlő szabályok feltárására. Az implementált algoritmusban a belső rétegben lévő neuronok száma megegyezik a kimeneti réteg neuronjainak számával. A kimeneti rétegben lévő neuronokkal a szavak szubjektív koordinátáinak lehetséges értékeit modelleztem. A

feladat specializációja alapján a szubjektív koordináták értékkészlete fixen rögzített. A neuronok közötti kapcsolatok súlyértékei [-1.0, 1.0] tartományba tartozó valós számok[10].

3.5. Kérdés- és válaszgeneráló modul

A modul feladata a kérdésként kiemelt mondatok, a mondatokból kérdésként kiemelésre kerülő szavak, valamint a feltett kérdésekre adható lehetséges válaszok meghatározása. A kérdés- és válaszgenerálást végző modul rendszertervét a *4. ábra* szemlélteti.



4. ábra: Kérdés- és válaszgeneráló modul

A felhasználónak először ki kell választani a dokumentumnak a kérdésként szereplő mondatait. Ez a modul bemeneti információként kapja a dokumentum mondatait annotáltan tartalmazó szöveges fájlt. Ezt követően a szakértői döntés alapján határozhatók meg azok a mondattípusok, melyekből kérdéseket lehet előállítani. Minden lényeges mondattípus külön annotációval lett ellátva a dokumentumban. A mondatok típusán kívül, a kérdésként szereplő mondatok számának beállítása is itt határozható meg (10-100 mondat). Ez szintén szakértői döntés alapján történik.

A mondatok meghatározását követően a kérdésként kiemelhető szavak kiválasztása következik. Ehhez iránymutatásként szolgálnak a szavaknak az osztályozási modultól kapott „kérdésként kiemelésre kerüljön-e" nevű szubjektív koordinátái. Mivel ezek a koordináták minden szót egyedileg jellemeznek, ezért előfordulhat, hogy egy mondaton belül több szó is megjelölhető kérdésként kiemelhetőnek. Ebben az esetben, a modul a lehetséges alternatívákat a neurális hálózat kimeneti neuronjainak bemeneti függvénye

[10] Bednarik et al., 2012

szerint rangsorolja. Azok a szavak, amelyeknél ez az érték magasabb, nagyobb valószínűséggel kerülnek kérdésként kiemelésre. A kérdésgenerálás utolsó lépéseként meg kell határozni a válaszként adható lehetséges alternatívákat.

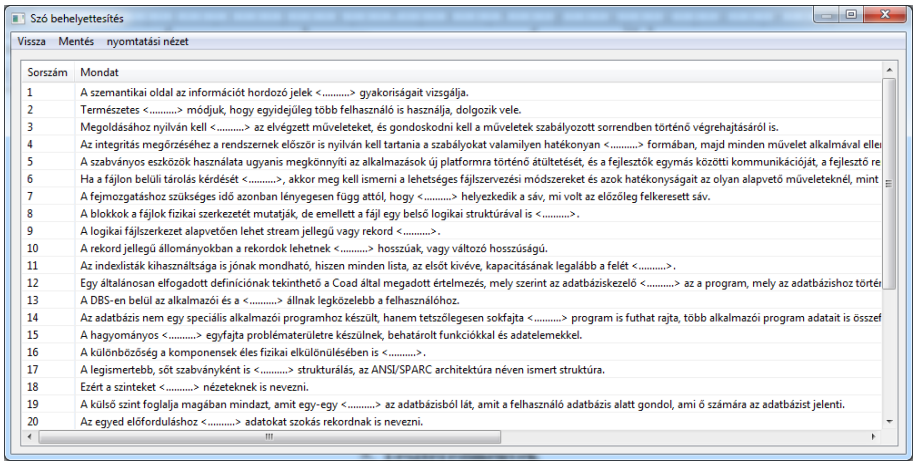
A válaszlehetőségeket előállító modul továbbfejlesztése az alapkoncepcióhoz képest három ponton valósult meg:

- minden válaszalternatíva szótövezett alakban jelenik meg,
- a kérdésként kiemelésre került szóval fogalmi szinten azonos szó is szerepeljen a lehetséges válaszalternatívák között,
- a fogalmi szinten azonos szó szófaja egyezzen meg a kérdésként kiemelt szó szófajával.

A szavak fogalmi szintű távolságainak meghatározására kidolgoztam egy háromszintű fogalomhierarchia modellt. Kategóriaszó-szófaj-szó szinteken megvalósítottam a dokumentum szavai közötti távolságokat. A fogalom-hierarchia legfelső szintjén a kategóriaszó található, melyben két tudományterülethez kerültek a szavak besorolásra: természettudomány és nyelvtudomány. A hierarchia középső szintjén helyezkednek el a szavak szófajai és a legalsó szinten pedig a kérdések előállításához használt szavak. Válaszalternatívákat tartalmazó fogalomszótár felépítése: fogalomhierarchiából kinyert referencia szó, referencia szóval fogalmi szinten lévő szó, referencia szót tartalmazó klasztertől egy távolságegységre lévő klaszterből kiválasztott szó, referencia szót tartalmazó klasztertől két távolságegységre lévő klaszterből kiválasztott szó és a referencia szót tartalmazó klasztertől három távolságegységre lévő klaszterből kiválasztott szó. Az előállított válaszalternatívákat az algoritmus összekevert sorrendben jeleníti meg a felhasználó előtt.

4. AQG grafikus felülete

Az AQG rendszer működését egy objektum-orientált nyelven implementált szoftverrel készítettem el. A kérdésgeneráló program egyik bemenete a dokumentumfájl, a másik a szótövezett fájl, amely definiálja a szavak szófaját, valamint szótövét és a harmadik fájl pedig a fogalomszótár. Az alkalmazás elindítása után a program generálja a feladatlapot, amely két változatban jeleníthető meg. Az elektronikus változat esetén mind az előállított kérdések, mind pedig a rájuk adható válaszok számítógépes környezetben jelenik meg a felhasználó előtt. Ebben a változatban a felhasználó a megválaszolandó kérdést tartalmazó mondatra kattintva tudja előhívni a kérdésre adható lehetséges válaszokat tartalmazó menüt. A válasz megadását követően a kiválasztott alternatíva automatikusan behelyettesítésre kerül a mondatba. A kitöltött feladatlap fájl formátumban elmenthető (.txt, .doc, .csv). A tesztlap nyomtatható formában való reprezentálásához a szoftver a kérdésként kivett szó helyét kipontozva jelzi a felhasználó számára, illetve a lehetséges válaszokat a mondatok alatt tűnteti fel egymás mellett felsorolt alakban. A tesztlap nyomtatott formátumának kitöltéséhez a felhasználónak a megfelelőnek ítélt szót aláhúzással, vagy a kipontozott részbe való beírásával kell jeleznie. A tesztlap elektronikus kitöltési formájára mutat egy példát az *5. ábra*.



5. ábra: Az automatikusan előállított tesztlap elektronikus változata

5. Teszteredmények

Az automatizált kérdésgeneráló rendszer eredményeinek tesztelésére az Adatbázis rendszerek című tantárgy jegyzete szolgált[11]. A felmérésben 40 hallgató vett részt, akik közül 50% tanulta a tantárgyat, a másik 50% nem ismerte annak tartalmát. Két feladatlapot kellett kitölteni, melyre 30 perc állt rendelkezésre. Az egyik a kérdésgeneráló rendszer által generált kérdéseket tartalmazta, a másik az oktató által manuálisan összeállított kérdésekből állt. Mindkét esetben a feleletválasztásos kérdésekként használt mondatok a tantárgy teljes írásos tananyagában annotált 500 mondat közül véletlenszerűen kerültek kiválasztásra. A tesztlapok 30 kérdést tartalmaztak, melyhez kérdésenként öt lehetséges választ rendelt a számítógép, illetve az oktató. Az eredmények kiértékelésre során a rendszerrel előállított feladatlapokon a 30 kérdésből 16 esetben természettudományhoz tartozó szó kiemelve kérdésként. A többi 14 mondat esetén a kiemelt szó a nyelvészettudományhoz tartozott. Ezzel szemben a manuálisan előállított feladatlapon a 30 kérdésből 23 esetén tartozott a kivett szó a nyelvészettudományhoz és 7 esetben a természettudományhoz.

A tantárgyat tanuló és a nem tanuló hallgatók által elért eredményeit vizsgálva mind a két csoport esetén a manuálisan készített kérdések átlageredménye jobb, mint az automatikusan generált kérdések eredményei. Az eredményeket a 1. táblázat foglalja össze.

[11] Kovács, 2004

1. táblázat: Automatikus és manuálisan előállított kérdések átlageredményei

Csoportok/Kérdések	Automatikusan előállított kérdésekre adott helyes válaszok átlaga	Manuálisan előállított kérdésekre adott helyes válaszok átlaga
Nem tanulta a tantárgyat	4,55	11,65
Tanulta a tantárgyat	18,66	21,3

6. Következtetések

A megvalósított szemi-automatizált kérdésgeneráló rendszer kidolgozásánál több fontos követelményt is teljesítettem. Ezek közül kiemelem a saját fejlesztésű rendszermodellt, ezen belül a modulokat, algoritmusokat és azok optimalizálását, melynek kifejlesztésével nem csak a tárigény csökkentést, hanem az előállított kérdés- és válaszalternatívák létrehozásánál sebességnövekedést is jelent. A kérdésgeneráláshoz magyar nyelven írt annotált dokumentumra van szükség, melyben eltérő annotációval van ellátva a fogalmak, definíciók, illetve kijelentő mondatok. A feladat elvégzéséhez szükség van nyelvtani elemzést végző alkalmazásra, mely megállapítja a dokumentum szavainak alapvető nyelvészeti tulajdonságait és egy háromrétegű fogalomhierarchiára, mely leírja a dokumentum szavainak szófajhoz, illetve kategóriaszóhoz tartozását.

A kérdésgeneráló rendszer előfeldolgozó moduljának teljes automatizálását a szöveges dokumentum DITA XML dokumentumformátumba való konvertálással valósítható meg. Ez további kutatási feladatot jelent számomra.

Irodalomjegyzék

Fajszi, B. – Cser, L. 2004. *Üzleti tudás az adatok mélyén*. Budapesti Műszaki és Gazdaságtudományi Egyetem.

Miller, G. 1995. *WordNet, a lexical database for English.* Communication of the ACM, Vol. 38, pp. 39–41.

Sumita, E. – Sugaya, F. – Yamamoto, S. 2004. *Automatic Generation Method of a Fill-in-the-blank Question for Measuring English Proficiency*, Technical report of IEICE, 104 (503), pp. 17–22.

Nielsen, R. 2008. *Question Generation, Proposed challenge tasks and their evaluation.* Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge. NSF, Arlington.

Gütl, C. – Lankmayr, K. – Weinhofer, J. – Höfler, M. 2011. *Enhanced Automatic Question Creator – EAQC* The Electronic Journal of e-Learning Vol. 9 Issue 1, pp. 23–38.

Németh, L. 2003. *A Szószablya fejlesztés*. pp. 3–4.

Tikk, D. 2007. *Szövegbányászat,* Typotex, Budapest.

Bodon, F. 2010. *Adatbányászati algoritmusok.* Free Software Foundation által kiadott GNU Free Documentation license 1.2-es, Budapest.

Bednarik, L. – Kovács, L. 2012. *Osztályozási feladatok a kérdésgenerálási mintarendszerben.* A Gépipari Tudományos Egyesület Műszaki Folyóirata (GÉP), LXIII. évfolyam.

Bednarik, L. 2012. *Automatizált kérdésgenerálás annotált szövegből*. Hatvany József Informatikai Tudományok Doktori Iskola, Miskolc.

Kovács, L. 2004. *Adatbázis rendszerek I.* Munkapéldány, http://www.iit.uni-miskolc.hu/iitweb/opencms/department/labs/iit-szolgaltatasok/www-db/Tantargyak/AB1/