MISKOLCI EGYETEM

GÉPÉSZMÉRNÖKI ÉS INFORMATIKAI KAR

AUTOMATIZÁLT KÉRDÉSGENERÁLÁS ANNOTÁLT SZÖVEGBŐL

Ph.D. értekezés tézise

KÉSZÍTETTE:

Bednarik László

okleveles mérnök-informatikus

AKI DOKTORI FOKOZAT ELNYERÉSÉRE PÁLYÁZIK

HATVANY JÓZSEF INFORMATIKAI TUDOMÁNYOK DOKTORI ISKOLA
ALKALMAZOTT SZÁMÍTÁSTUDOMÁNY TÉMATERÜLET
ADAT- ÉS TUDÁSBÁZISOK, TUDÁSINTENZÍV RENDSZEREK TÉMACSOPORT

Miskolc, 2012

Bednarik László

AUTOMATIZÁLT KÉRDÉSGENERÁLÁS ANNOTÁLT SZÖVEGBŐL

Ph.D. értekezés tézise

Témavezető:
Dr. habil. Kovács László
egyetemi docens

Társtémavezető:
Prof. Dr. Juhász Imre
egyetemi tanár

Miskolc, 2012

VÉDÉSI BIZOTTSÁG

Elnök:

Prof. Dr. Tóth Tibor, DSc	ME, egyetemi tanár

Tartalék elnök:

Prof. Dr. Szigeti Jenő, CSc	ME, egyetemi tanár

Titkár:

Dr. Mileff Péter, PhD	ME, egyetemi adjunktus

Tagok:

Dr. Bujdosó Gyöngyi, PhD	DE, egyetemi adjunktus
Dr. Dudás László, CSc	ME, egyetemi docens
Dr. Kovács Emőd, PhD	EKF, főiskolai docens
Dr. Kovács Szilveszter, PhD	ME, egyetemi docens
Dr. Kusper Gábor, PhD	EKF, főiskolai docens

Opponensek:

Dr. Johanyák Zsolt Csaba, PhD	GAMF Kar, főiskolai tanár
Dr. Sasvári Péter, PhD	ME, egyetemi docens

TARTALOMJEGYZÉK

1. BEVEZETÉS

A Ph.D. munkám keretében végzett kutatásom fő célja egy szemi-automatizált kérdésgeneráló mintarendszer megtervezése és megvalósítása volt. A mintarendszer feladata annotált szöveges dokumentumból kérdések és válaszlehetőségek előállítása.

A szövegbányászat, mely az 1990-es évek végén jelenik meg, a számítástudomány szöveges dokumentumok feldolgozásával és elemzésével foglalkozó szakterülete [5]. Kialakulása a digitalizált vagy nyomtatott elektronikus dokumentumok és médiaanyagok számának növekedésével indult el, amit az Internet megjelenése és térhódítása tovább fokozott. A fejlődés hatására jelentősen megnőtt az elektronikus szöveges tartalmak mennyisége, amelynek nemcsak tárolása okoz problémát, de hatékony feldolgozásának, rendszerezésének kérdése is egyre inkább előtérbe kerül. A szöveges dokumentum speciális adathalmaz, mely általában strukturálatlan formában áll rendelkezésre. Ugyanakkor gazdag információtartalommal rendelkezik. A szövegbányászat fontos eszköze az informatikával támogatott oktatási eszközök fejlődésének, mely a szöveges dokumentumokból kinyerhető szabályok feltárásával foglalkozik [10]. Az elektronikus oktató rendszerek egyik fontos jövőbeli irányát jelentik az adaptív és szemantika-orientált számítógépes oktató alkalmazások. Napjainkban a kutatások egyik fő jellemzője, hogy a hagyományos oktatási keretrendszerek mellett megjelennek a témakör ismeretanyagát tároló tudásbázisok, szemantikai adatbázisok. Ezen adatbázisok tudásanyagára építve rugalmasabban és funkcionálisan gazdagabban kezelhető a tananyag. A rugalmasság elsődlegesen a hallgatói igényekhez történő alkalmazkodás képességét jelenti. Az adaptivitás a rendszer több funkciójában is megjelenik.

A hallgatók tudásának ellenőrzése, a számonkérés a szemantika-orientált oktatási keretrendszerek egyik fő funkciója. A kérdések megalkotása önmagában is összetett feladat, hiszen illeszkednie kell a tématerülethez, az ellenőrzés céljához, a hallgató aktuális állapotához [2].

A kutatás abból az előfeltevésből indul ki, hogy a létező megközelítések és eszközök nem kielégítően rugalmasak és kiterjeszthetők ahhoz, hogy támogassák a széles körben alkalmazott tankönyveket [16], [17] és tanulási körülményeket. Az egyik nehézség, hogy ezek a módszerek specifikus, domináns nyelvekre korlátozottak az alkalmazott nyelvtani modulok miatt.

Másik probléma, hogy az általuk használt szemantikai tudásbázisok csak az adott nyelven használhatók.

A feladat mérete és bonyolultsága miatt az értekezés csak a feleletválasztós, kiegészítő kérdéstípusok automatikus előállításával foglalkozik. Az eredmények tesztelése során a kérdésgenerálás alapjául szolgáló dokumentumokat a műszaki felsőoktatásban használt tantárgyak tananyagai alkották. A kérdésgenerálás külső szemantikai adatbázis felhasználása nélkül került megvalósításra. Munkám elsődleges célja rugalmas és nyitott keretrendszer megtervezése, implementálása és gyakorlati működésének igazolása, amely a magyar nyelvet használja a kérdés- és válaszlehetőségek előállítására.

1.1. Irodalmi áttekintés

Az automatizált kérdésgenerálás alapfeltétele a témakör fogalmainak szemantikai megismerése. George Miller és munkatársai a Princeton Egyetem Kognitív Tudományi Laboratóriumában az angol nyelv szavai és fogalmai köré szerveződő lexikális szemantikai hálózatot készítettek [Miller, 1990]. Ez a WordNet-nek nevezett szinkron nyelvi tudást reprezentáló speciális szemantikai hálózat. A WordNet olyan elektronikus lexikális szemantikai adatbázis, melyben a nyelvi fogalmak hálózatba szerveződnek. A fogalmakat szinonimahalmazok (synsetek), a közöttük lévő kapcsolatokat szemantikai relációk (hipernima, meronima, antonima stb.) reprezentálják. A WordNet-en elérhető információk alapján hatféle kérdés hozható létre: definíció, szinonim, antonim, hiperném, hiponém és feleletválasztós kérdések. Ahhoz, hogy kinyerjünk valamilyen adatot a WordNet-ből, ki kell választani az adott szónak a megfelelő jelentését. Adott szóhoz hozzá van rendelve több szó is a WordNet rendszerben. Leggyakrabban a beviteli rész csak az adott szót tartalmazza beszéd részeként [Brown, 2004]. A definiáló kérdés a szóra vonatkozó definíciót kínálja fel. Ez elérhető a WordNet jegyzet részből. A szinonima típusú kérdésekben a kérdéses szót összekapcsolja a rokon értelmű szavával és a rendszer kinyeri a WordNet-ből az adott szónak a rokon értelmű szavait. Az antonim kérdésektől azt várják el, hogy a szót párosítsa összes az ellentétes szavával. A hiperném és hiponém kérdéstípusok hasonló szerkezetűek. A hiperném a fogalmak egy általános meghatározás osztályába sorolását jelenti. A feleletválasztós kérdés típusnál adott a fő kérdés és ezt követi számos válasz, amelyekből csak egy a megfelelő.

A kérdésgeneráló keretrendszer moduljainak meghatározásakor az egyik legkorábbi alapmodell, a Coniam [Coniam, 1997] által kifejlesztett mintarendszer szolgált alapul. A javasolt architektúra első modulja a forrás dokumentumok szövegének strukturális tagolását, előfeldolgozását végzi. Ezt követően a kérdés tárgyát hordozó mondatok kijelölése következik, melynek végrehajtásához a támogatott tanulási módszereket alkalmazták. A következő lépésben a mondaton belül a kiemelt szó kerül meghatározásra. A harmadik fázisban a választási listába kerülő szavak kollekciója épül fel egy külső adatforrás, szemantikai adatbázis alapján [Collins, 2007]. Sumita [Sumita, 2004] is javasolt egy hasonló elvű automatizált generáló módszert a feleletválasztós kérdések előállítására. Ebben a módszerben a mondatokban szereplő ige került kiválasztásra a kérdés előállításában. Munkájában a modul három fő lépésből áll: feldolgozandó mondat kiválasztása, meg kell határozni a mondat üres részét és generálni kell a hiányos mondatokat. A mondatok kiválasztása, az üres helyek és a hiányos mondatok meghatározása a gépi tanulási módszerek segítségével történik statisztikai és diszkriminatív modellek felhasználásával. Rus és szerzőtársai [Rus, 2007] minták, sablonok és speciális jelölő nyelv segítségével vezettek be egyedi módszereket a kérdésgeneráláshoz. A mintákat szemantikai, lexikális és szintaktikai struktúrák jellemzik, míg a sablonok módszereket írnak le a kérdésgeneráláshoz.

Az automatizált kérdésgeneráló rendszerek fejlesztésével foglalkozó kutatások eredményei 2008 után már komplex architektúrák formájában jelentek meg. Nielsen és szerzőtársai [Nielsen, 2008] megalkottak egy olyan modellt, amelyben a kérdések automatikus előállítása a következő lépésekből áll. Elsőként a forrás dokumentumból ki kell választani azt a tartalmat, amire szeretnénk, hogy az előállítandó kérdés vonatkozzon. Ezt követően az elvárt válasz ismeretében ki kell választani a kérdés típusát. Végül elő kell állítani azt a kérdést, amely megfelel a megadott tartalom, illetve kérdéstípus paramétereinek [Piwek, 2008].

Az OpenLearn olyan nyílt keretrendszer, amely XML formátumot használ az oktatási anyag tárolásához [8], [12]. A CEIST módszer [Piwek, 2009], amit az OpenLearn nyílt keretrendszeréhez fejlesztettek ki, rendelkezik egy előfeldolgozó fázissal, ahol a dokumentumtartalmat tiszta szövegformátumból átkonvertálják egy szövegelemző fastruktúrába a nyelvtani elemző modul segítségével. Az egység fő motorja a mintaegyeztető algoritmus, amit ahhoz használnak, hogy megtalálják az előzőleg egyeztetett mintákat a fában. A fában

megtalált mondatokhoz konverziós szabály sémáját felhasználva generálnak kérdéseket [Piwek, 2009].

Gütl és munkatársai [Gütl, 2008] 2008-ban elkészítettek egy automatizált kérdésgeneráló rendszert (AQC). A kifejlesztett prototípus tapasztalatain és eredményein alapulva 2010-ben továbbfejlesztették a rendszert (EAQC). A továbbfejlesztett rendszer több különböző nyelv tanulási tananyagából is támogatja automatikus tesztkérdések készítését. A rugalmas tervezés lehetővé tette, hogy önállóan használható legyen az eszköz és adaptálni lehessen speciális tanulási környezetbe. Az EAQC rendszer három fő modulból áll [Gütl, 2011]:

- Az előfeldolgozási modul feladata a különböző bemeneti fájlformátumok átalakítása, nyelvfelismerés és átalakítás belső XML sémává, amely tartalmazza az összes szükséges adatokat a további feldolgozáshoz. A jelenlegi modell az angol [WordNet, 2010] és a német nyelvet [GermaNet, 2009] támogatja, de lehetőség van más nyelvek és eszközök beépítésére is a modulba.
- A fogalmak kinyerésére szolgáló modul strukturális, statisztikai és szemantikai elemzést végez és kinyeri a legmegfelelőbb fogalmakat a dokumentumból.
- A kérdésgeneráló modulban történik a kérdéstípusok kiválasztása: nyitott végű mondatok, egyválasztós, többválasztós és feleletválasztós kérdések.

Az értekezésben kidolgozott kérdésgeneráló mintarendszer esetén, olyan rendszermodellt terveztem és valósítottam meg, amely működőképes külső szemantikai adatbázis nélkül is, hiszen WordNet jellegű adatforrás nem állt rendelkezésre a kiválasztott területen és nyelvben. Emiatt a fogalomszótárt belső saját fejlesztésű modullal valósítottam meg. További eltérés a szakirodalomban meglévő kérdésgeneráló modellekhez képest az 1.2. fejezet tartalmazza.

1.2. A kutatás célja

Az értekezés mintegy határterületként az oktatási, szövegbányászati és tudásmérnökségi diszciplínákat foglalja magába. A kutatás célja olyan rendszermodell megtervezése és megvalósítása, amely képes annotált magyar nyelvű szöveges dokumentumból, megadható mondattípusok alapján, feleletválasztós, kiegészítő kérdések automatikus előállítására.

A rendszermodell az alábbi fő funkcionális modulokat foglalja magába:

- **előfeldolgozás**, feladata az eltérő szerkezetű és formátumú dokumentumoknak – egységesítési, formalizálási és normalizálási eljárások alkalmazásával – a kérdésgenerálás számára egységesen kezelhető és lényegkiemelést támogató strukturált formára alakítása;
- **klaszterezés**, feladata a dokumentum szavainak szeparált csoportokba szervezése, ahol az azonos csoportba sorolt szavak – a definiált kritériumok értelmében – egymáshoz minél hasonlóbbak legyenek, míg a különböző csoportba soroltak egymástól minél különbözőbbek legyenek;
- **osztályozás**, feladata olyan szabályrendszer feltárása, mely képes az egzaktul meghatározható objektív koordinátákkal definiált térben adott szavakat automatikusan elhelyezni az emberi megérzésekre alapuló szubjektív koordinátákkal definiált térben,
- **kérdésgenerálás**, feladata a bemenetként kapott dokumentumból a kérdésként szolgáló mondatok meghatározása, a mondatokból a kérdést reprezentáló szó kiemelése, valamint a kérdésre adható lehetséges válaszok automatikus előállítása.

A kidolgozott mintarendszer az alábbi tulajdonságokkal rendelkezik:

- támogatja a különböző bemeneti fájlformátumokat a helyi fájlrendszerből és az Internet erőforrásokból,
- magyar nyelvű támogatás,
- a tudásterülettől és dokumentumstruktúrától való függetlenség,
- tesztkérdések készítése annotált szövegből,
- válaszlehetőségek előállítása belső, saját fejlesztésű szemantikai adatbázisból,
- annotáció támogatása,
- feleletválasztós, kiegészítő feladatok támogatása,
- elektronikus és nyomtatható feladatlapok készítése,
- konfigurálhatóság, modularitás és kiterjeszthetőség,
- együttműködési képesség a létező e-learning rendszerekkel [15], [16], [17].

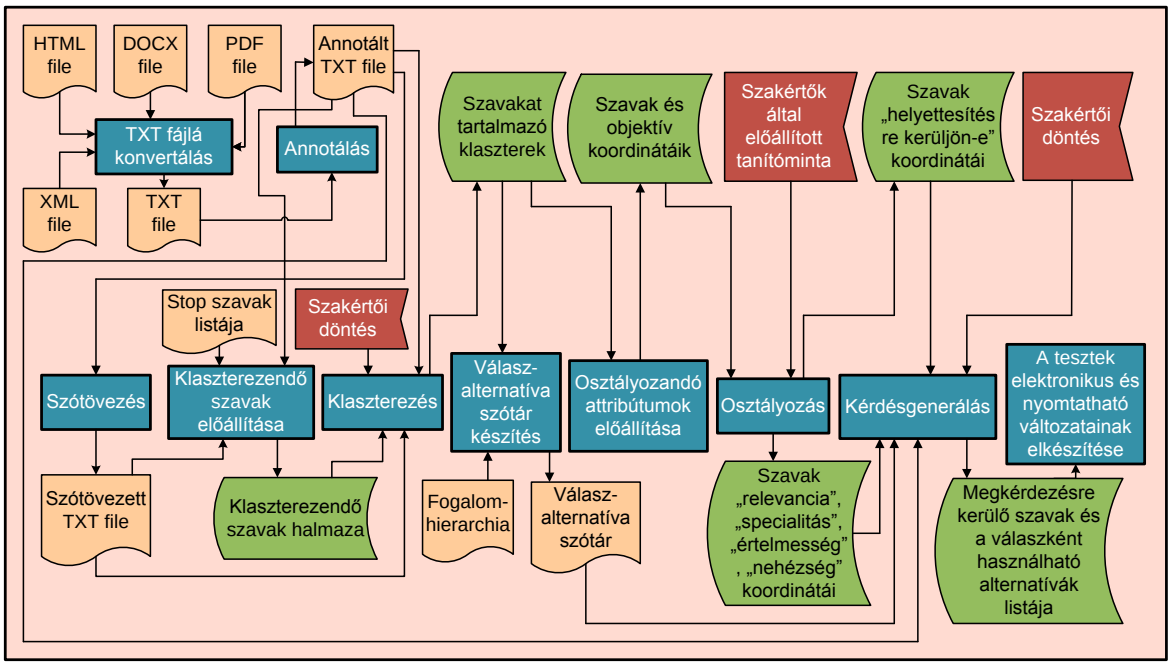
A jelenleg fellelhető automatizált kérdésgeneráló rendszerek jelentős része idegen nyelvű szövegek kezelésére fejlesztették ki. A magyar nyelvűek kísérleti stádiuma, illetve a felsorolt jellemzőknek csak részleges támogatása miatt

szükséges volt saját fejlesztésű, gyakorlati életben alkalmazható rendszer kidolgozására.

2. Új tudományos eredmények

2.1. Az automatizált kérdésgenerálás rendszertervének kidolgozása

Az automatizált kérdésgenerálás megvalósítására olyan rendszermodellt dolgoztam ki, melyben több egymással adatkapcsolatban álló alrendszer együttes összehangolt működése biztosítja a feladat megoldását. Annak érdekében, hogy a rendszer egyes moduljai a későbbi fejlesztések során egymástól függetlenül elemezhetők és módosíthatók legyenek minden modul számára bemeneti, illetve kimeneti interfészeket definiáltam. A bemeneti interfészek írják le azokat az adatokat, amelyeket a modulok igényelnek az általuk reprezentált feladat elvégzéséhez. A kimeneti interfészek azokat az adatokat definiálják, amelyeket a modulok szolgáltatnak a hozzájuk közvetlenül kapcsolódó további modulok számára. A disszertációban megtervezett automatizált kérdésgenerálás feladatát ellátó rendszermodell funkcionális rendszertervét a 2.1. ábra szemlélteti [14].



2.1. ábra: A rendszermodell funkcionális rendszerterve

Ez alapján jól követhető, hogy milyen adatokra és tevékenységekre van szükség a bemenetként kapott szöveges dokumentum alapján végzett automatizált kérdés- valamint válaszgenerálás megvalósításához. A kidolgozásra került főbb modulok és feladataik az alábbiak:

1. *Szöveges dokumentumok előfeldolgozását végző modul*

   A modul feladata a kérdésgenerálás alapjául szolgáló különböző formátumú dokumentumok egységes formára hozatala, amelynek során a formázási, strukturális és szerkesztési adatok mellőzésével pusztán a szöveges információk kerülnek kinyerésre. A konverzió célja a további modulok számára egységesen kezelhető a feladat szempontjából csak a releváns információkat tartalmazó szöveges dokumentum biztosítása. Az előfeldolgozást végző modul támogatja a bemenetként szolgáló nagyméretű szöveges dokumentum mondatainak annotációval megvalósított szűrését, illetve a mondatok jól definiált kategóriákba sorolását (fogalom, definíció, kijelentő mondat). Ezek alapján a modul már csak a szűrésben beállított kategóriákhoz tartozó mondatokat teszi elérhetővé a kérdésgenerálás további moduljai számára.

2. *Szótövezést végző modul*

   A modul feladata a dokumentumban szereplő szavak alapvető nyelvtani elemzése. Ezt a feladatot a Szószablya keretrendszer interneten, szabadon használható alkalmazással valósítottam meg. Ennél a modulnál fő feladatom egyrészt az előző modulok által szolgáltatott információknak a Szószablya keretrendszerhez való illesztése, másrészt a szótövező keretrendszer által szolgáltatott információknak a kérdésgenerálás további moduljaihoz való illesztése jelentette. A szótövezéssel megszerzett releváns információk: szótő, szófaj, ragozottság, szótagszám.

3. *Klaszterezést végző modul*

   A modulnak feladata a nagyszámú szóból álló dokumentumok vizsgálandó egységeinek lecsökkentése a dokumentum bizonyos szempontból egymáshoz közeli szavainak klaszterekbe rendezésével. Ezáltal a legkisebb vizsgálható egységet a szavak helyett a klaszterek reprezentálják, amely lényeges erőforrás-megtakarítást jelent. A klaszterezést elvégeztem a szavak közös előfordulási gyakoriságára épülő, valamint a szavak objektív koordinátái által definiált többdimenziós térben vizsgált vektortér alapú távolságra épülő stratégiák alapján is.

4. *Osztályozást végző modul*
   A modul feladata a dokumentum szavainak objektíven mérhető koordinátái alapján a szavakhoz definiált szubjektív koordináták meghatározása. A szubjektív koordináták egyrészt a kérdésként kiemelésre kerülő szó meghatározásában nyújtanak erős támogatást, másrészt a kiemelt szóhoz (mint helyes megoldáshoz) szubjektív szempontból definiált szinonimák megtalálására is használhatóak.
5. *Kérdésgenerálást végző modul*
   A modul feladata az előző modulok által szolgáltatott információk alapján a dokumentum mondataiból kiválasztani a kérdésekre használhatókat és meghatározni a kérdésként kiemelendő szót. A modul további feladata a kiemelendő szóhoz megfelelő válaszalternatívák előállítása. Ehhez a szavak klaszterezésével nyert távolságadatok, valamint a saját definiálású fogalomhierarchia szolgál alapul.

Az automatizált kérdésgenerálást végző rendszer számára definiáltam azokat a tényezőket, melyek meghatározzák a módszer alkalmazhatósági területeit. Ezek alapján a kérdésgeneráláshoz magyar nyelven írt annotált dokumentumra van szükség, melyben eltérő annotációval vannak ellátva a fogalmak, definíciók, illetve kijelentő mondatok. A feladat elvégzéséhez szükség van nyelvtani elemzést végző alkalmazásra, mely megállapítja a dokumentum szavainak alapvető nyelvészeti tulajdonságait. Ezeken felül szükség van egy háromrétegű fogalomhierarchiára, mely leírja a dokumentum szavainak szófajhoz, illetve kategóriaszóhoz tartozását.

1. tézis:
Meghatároztam a rendszer adatmodelljének- és funkciómodelljének elemeit. Definiáltam az automatizált kérdésgenerálás megvalósításához szükséges feladatokat, a feladatokhoz kapcsolódó bemenő és kimenő adatokat. Korlátfeltételek meghatározásával kijelöltem a módszer alkalmazhatósági területeit. Kidolgoztam a részfeladatok összehangolt ellátására szolgáló moduláris rendszertervet és egy dokumentum szerkezeti modellt. Bemutattam a kidolgozott modell elvi működőképességét [3], [14].

2.2. Klaszterezési algoritmusok kidolgozása és optimalizálása a kérdésgeneráláshoz szükséges objektív koordináták meghatározásához

A szöveges dokumentumok alapján végzett kérdések, illetve a kérdésekre adható válaszok automatikus előállításához a dokumentum szavainak több szempont alapján végzett elemzésére van szükség. Az elemzésekkel két fő célt teljesítettem. Egyrészt meghatároztam a szavaknak azokat a jellemzőit, amelyek a kérdésként való kiemelésük alapjául használhatóak. Másrészt feltártam a szavaknak az egymáshoz való viszonyát, amit a kérdésre adható válaszok előállítására alkalmaztam. Tesztekkel igazoltam, hogy a feladatom során kidolgozott klaszterezési algoritmusok lehetővé teszik a gyakorlati életben szükséges több 100 oldalas dokumentumok kezelhetőségét is.

A kérdésként kiemelt szó helyettesítésére felkínált szavakkal szemben az egyik leglényegesebb követelmény, hogy mérni lehessen általuk a hallgatók tudásszintjét. Azaz, a felkínálásra került alternatívák a dokumentum szavai által definiált teret a kérdésként kiemelésre került szóhoz (helyes válasz) való távolságuk alapján egyenletesen fedjék le. Ennek egzakt meghatározásához ismerni kell a dokumentum minden szavának az összes többi szóhoz való távolságát. Ez azonban a gyakorlati életben megkívánt méretű feladatok esetén a rendelkezésre álló számítási kapacitásokkal nem megoldható. A klaszterezés révén a szavak közötti hasonlóságokat a szavakat tartalmazó klaszterek közötti távolságokkal definiáltam. Ezáltal a klaszterekbe kerülő szavak átlagos számával arányos sebességnövelést értem el. Annak érdekében, hogy a kidolgozott kérdés, illetve válaszgenerálási koncepciót különböző igények esetén is alkalmazhatóvá tegyem a klaszterezést két eltérő távolságképzési koncepció alapján is megvalósítottam. Az első módszerben két szó hasonlóságát a dokumentum azon mondatainak a számával definiáltam, melyekben mindkét szó közösen előfordul (2.1) [7]:

$$S_{i,j} = f_{i,j} \,/\, \max(f_i,\ f_j) \tag{2.1}$$

ahol: S_{ij} az i. és j. szó távolságviszonyát reprezentáló [0, 1] intervallumba eső számérték (a nagyobb érték a szavak közötti kisebb távolságot jelöli); 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 szótávolságra épülő távolság-meghatározási koncepciót a QTC (Quality Treshold Clustering) algoritmus feladatspecifikus adaptálásával valósítottam meg, melyben a klasztereket előre definiált egyenlő sugarú körökkel modelleztem. Az algoritmus a klaszterek átmérőjét a következőképpen definiálja (2.2) [1]:

$$d = max_{i,j}\left\{\sqrt{\left(x_i - x_j\right)^2}\right\} \tag{2.2}$$

ahol: d a klaszter átmérő, x_i az i-dik pont összes objektív koordinátája, x_j a j-dik pont összes objektív koordinátája. A klaszterezést végző optimalizációs algoritmus célfüggvényét a klaszterszám minimalizálásával, a korlátfeltételét pedig a dokumentum minden szavának legalább egy klaszterhez tartozásával definiáltam. Az algoritmusokat implementáltam a szélességi, mélységi valamint a legjobbat először keresési stratégiák alkalmazásával.

A másik koncepcióban a szavak távolságát a szavak objektív módszerekkel meghatározható koordinátái által definiált többdimenziós térben mérhető vektortér alapú távolságuk alapján határoztam meg. A szavaknak a nyelvészeti, valamint statisztikai módszerekkel objektíven mérhető tulajdonságait a szavak *objektív koordinátáinak* nevezzük. Az alkalmazott objektív koordináták: szófaj, ragozottság, dokumentumon belüli előfordulási szám, mondaton belüli előfordulási szám, mondaton belüli hely.

Az objektív koordinátákkal számoló koncepciót a szakirodalomban BIRCH [8] néven ismert klaszterezési módszer feladatspecifikus implementációjával valósítottam meg. A vektortér alapú távolságra épülő klaszterezéssel az volt a célom, hogy olyan klaszterezési lehetőséget is biztosítsak, amely nagyméretű dokumentumok esetén is elfogadható időn belül képes elvégezni a szavak klaszterezését. A küszöbérték és a maximális elágazási tényező a BIRCH algoritmus leglényegesebb paramétere. Kidolgoztam a BIRCH algoritmus költségigényét:

- egy pont elhelyezésének költségigénye,
- egy halmaz szétbontásának maximális költségigénye,
- BIRCH algoritmus számítási idejének költsége.

Meghatároztam a BIRCH algoritmus paramétereinek optimalizálás nélkül és optimalizálással kapott költségeit.

2. tézis:
Elemeztem a szavak távolságát mutató klaszterezési eljárások lehetőségeit és kidolgoztam egy alkalmas attribútum rendszert. Kidolgoztam és optimalizáltam egy paraméterezhető minőségi küszöbérték alapú klaszterező eljárást. Az automatizált kérdésgeneráló rendszerben a kijelölt klaszterek megfelelően reprezentálják a szavak szemantikai szerepkörét a vizsgált tartományon belül [1], [7], [8].

2.3. Osztályozási algoritmus kidolgozása és optimalizálása a kérdésgeneráláshoz szükséges szubjektív koordináták meghatározásához

A mondatokból kérdésként kiemelésre kerülő szó meghatározását objektív és szubjektív szempontok elemzésével végeztem el. Az objektív szempontokat a szavak nyelvtani elemzése, valamint a klaszterezéssel nyert információk alapján határoztam meg. A szubjektív szempontokat a szavak – előzetesen rögzített – szubjektív koordinátákkal meghatározott térben való elhelyezkedésével definiáltam. A *szubjektív koordináták* a szavaknak azokat az egzaktul nem meghatározható jellemzőit jelölik, melyek az embernek az adott szóra vonatkozó megítélését fejezik ki. A szubjektív tér dimenzióit a következő öt koordinátával írtam le: relevancia, specialitás, értelmesség, nehézség, „kérdésként kiemelésre kerüljön-e”. Az első négy koordináta értékkészletét öt különböző értékkel jellemeztem. Az egyes értékekkel a dokumentum szavainak az adott koordinátával kifejezett jellemzőhöz való viszonyát határoztam meg. Az ötödik szubjektív koordinátával a szakértő véleményét reprezentáltam az adott szó kiemelése szempontjából. Kutatásom során olyan osztályozó algoritmust dolgoztam ki, amely képes megteremteni a kapcsolatot a dokumentum szavainak egzaktul mérhető objektív koordinátái és az emberi megítélést kifejező szubjektív koordináták között. Az osztályozási feladat magját háromrétegű előrecsatolt neurális hálózattal valósítottam meg, mely egy belső réteget tartalmaz. A hálózatot mind strukturális, mind pedig működési szempontok alapján a konkrét feladat speciális tulajdonságaihoz illesztettem. A hálózat betanítását a felügyelt tanítási módszer alapján végeztem el. Az objektív és szubjektív koordináták közötti transzformáció szabályainak feltárását optimalizálási feladatként valósítottam meg, melynek során az osztályozó algoritmus optimalizálásának célfüggvénye a $\frac{helyesen\ osztályozott\ szavak\ száma}{osztályozott\ szavak\ száma}$ összefüggéssel felírt tudásszintjének maximálásaként definiáltam. Az

optimalizálás korlátfeltétele a tanulásra fordítható idő, illetve a maximális lépésszám betartása jelentette.

Mivel az objektív koordináták értékkészlete különböző dokumentumok, illetve klaszterezési paraméterek függvényében jelentősen eltérhet, ezért a bemeneti rétegben lévő neuronok számának pontos meghatározására külön algoritmust fejlesztettem ki. Ez az algoritmus az osztályozás megkezdése előtt feltárja az objektív koordináták által felvehető értékek halmazát és minden objektív koordináta minden lehetséges értékéhez külön neuront vesz fel a bemeneti rétegben. A kimeneti rétegben lévő neuronokkal a szavak szubjektív koordinátáinak lehetséges értékeit modelleztem. Mivel a szubjektív koordináták értékkészlete fixen rögzített (4 darab 5 értékű, valamint 1 darab 2 értékű), ezért a kimeneti neuronok számosságát 22 darabban állapítottam meg. A neuronok közötti kapcsolatok súlyértékei [-1.0, 1.0] tartományba tartozó valós számokkal modelleztem. A hálózat bemeneti, illetve kimeneti neuronjai közötti átviteli formulát a (2.3) alapján definiáltam:

$$a_i = g\left(\sum_{j=0}^{n} W_{j,i} * g\left(\sum_{k=0}^{m} W_{k,j} * a_k\right)\right) \tag{2.3}$$

ahol: a_k a k jelű bemeneti rétegbeli neuron aktiváltsági állapota; a_i az i jelű kimeneti rétegbeli neuron aktiváltsági állapota; m a bemeneti rétegben lévő neuronok száma; n a rejtett rétegben lévő neuronok száma; $W_{k,j}$ a k jelű bemeneti rétegbeli neuron és a j jelű rejtett rétegbeli neuron közötti kapcsolat erőssége; $W_{j,i}$ a j jelű rejtett rétegbeli neuron és az i jelű kimeneti rétegbeli neuron közötti kapcsolat erőssége; g aktivációs függvény, mely 1-es értéket reprezentál, ha a paramétereként kapott összeg pozitív, 0-át egyébként.

A neurális hálózat tanítása szakértő bevonását igénylő hosszadalmas feladat, mely során a kívánt tudásszint eléréséhez a dokumentum mondatainak legalább 1/3-a esetén meg kell határozni a mondat szavainak szubjektív koordinátáit. Annak érdekében, hogy a megszerzett tudás több dokumentum esetén is alkalmazható legyen a tanulás eredményeinek újrahasznosítására két módszert fejlesztettem ki [3]:

- neurális hálózat adatainak elmentése,
- osztályozás eredményéül szolgáló mondatokat elmentése.

3. tézis:
Meghatároztam a kérdésként kiemelésre kerülő szavak kiválasztásához szükséges bemenő objektív koordinátákat. Kidolgoztam egy neurális hálózaton alapuló osztályozó algoritmust, amely képes feltárni a dokumentum szavainak objektív, valamint szubjektív koordinátái közötti kapcsolatot. A kidolgozott CPN adaptációs hálózat teljesíti az automatizált kérdésgeneráló rendszerben elvárt pontosságot [3], [14].

2.4. Automatizált kérdés- és válaszgenerálási algoritmusok kidolgozása, az eredmények gyakorlati igazolása

A 2. és 3. tézisben ismertetett algoritmusok eredményeire építve kidolgoztam a kérdések és válaszok automatizált előállításának módszerét. Ennek első lépéseként a klaszterezett és osztályozott dokumentumhoz meghatároztam azokat a mennyiségi és minőségi kritériumokat, amelyek révén a kérdésként használt mondatok kiválasztásra kerülnek. A minőségre vonatkozó kritériumokat a dokumentumban szereplő mondatok típusai közötti választással definiáltam. Ehhez a mondatok típusait annotációval különböztettem meg egymástól. A mondatokból kérdésként kiemelésre kerülő szót az osztályozást végző algoritmus eredményei alapján határoztam meg. A válaszlehetőségeket előállító modell továbbfejlesztése három ponton valósul meg:

- minden válaszalternatíva szótövezett alakban jelenjen meg a válaszadó előtt,
- az alkalmazott statisztikai távolságmérés mellett „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 modelljét. Ebben *kategóriaszó-szófaj-szó* szinteken reprezentáltam a dokumentum szavai közötti távolságokat. A hierarchia legalsó szintjén helyezkednek el a kérdések előállításához használt dokumentumnak az elemzésből kizárásra nem kerülő szavai szótövezett alakban. A szavak feletti első kategorizálási rétegben a szavak szófaj szerinti besorolása valósul meg. Ez a réteg gondoskodik arról, hogy az azonos szülő csomóponthoz tartozó szavak szófaja megegyezzen. A fogalomhierarchia

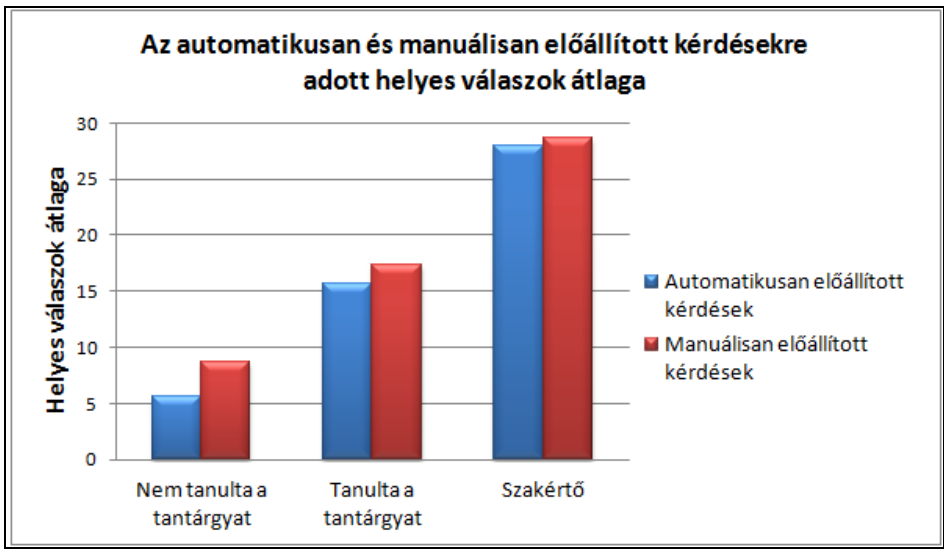
legfelső szintjén az azonos szófajhoz tartozó szavak a kategória szavak alapján kerülnek tovább kategorizálásra. A szó kategória szava határozza meg azt a fogalmi tématerületet, amelyhez a fa levél-csomópontjaiban lévő szavak szemantikai szempontból tartoznak.

A lehetséges válaszok meghatározásának első lépéseként meghatározásra került a kérdésként kivett szó szófaja. A kidolgozott modellben minden kérdésként kivett szó helyére öt lehetséges alternatívát kínál fel az algoritmus. A kérdésre adható legjobb választ természetesen az a szó képviseli, ami a mondatból kivételre került. Ezért ennek a szónak mindenképpen szerepelnie kell az automatikusan előállított öt lehetőség között. Abban az esetben, ha a mondatból kivett szó szótöve szerepel a fogalomhierarchiában, valamint van ezzel a szóval azonos szinten lévő más szó is ezen a hierarchiaszinten, akkor az algoritmus véletlenszerűen választ egy szót ezek közül. A fennmaradt három, illetve négy további alternatíva előállítása a dokumentum szavainak klaszterezésével nyert távolságok alapján történik. Az algoritmus meghatározza azt a klasztert, amely a mondatból kivett szót tartalmazó klasztertől a legtávolabb helyezkedik el. Ennek a klaszternek a kivett szót tartalmazó klasztertől való távolságát annyi egyenlő részre osztja fel, ahány alternatíva meghatározására még szükség van. Ezt követően az algoritmus sorra veszi azokat a klasztereket, amelyek a kivett szót tartalmazó klasztertől *n* távolságegységnyi távolságra vannak. Ezzel kerül meghatározásra az a távolságegység, amilyen lépésekkel halad az algoritmus a kivett szót tartalmazó klasztertől a hozzá legtávolabbi klaszterig. Az előállított válaszalternatívákat az algoritmus összekevert sorrendben tárja a kérdések megválaszolója elé. Ezzel a módszerrel ötvöztem az objektív és a szubjektív aspektusok nyújtotta lehetőségeket a szöveg lényegét minél inkább kifejező válaszok előállítása érdekében.

Az elkészült automatizált kérdésgeneráló mintarendszer tesztelésére és az eredmények igazolására a 2011/12 tanév második félévben került sor a Miskolci Egyetem Comenius Főiskolai Karán. A tesztelés célja a kidolgozott algoritmusok gyakorlati alkalmazhatóságának igazolása volt. Az elemzés kiterjedt az elkészített szoftverrel automatikusan előállított kérdések hallgatók által való elfogadhatóságának, illetve értelmezhetőségének elemzésére.

A felmérésben két feladatlapot kellett minden hallgatónak kitöltenie. Az egyik a kidolgozott kérdésgeneráló mintarendszer által generált kérdéseket tartalmazta, míg a másik a tantárgyat oktató személyek által manuálisan

összeállított kérdésekből állt. A helyes válaszok átlagát mindhárom csoportba tartozó hallgatók esetén a 2.2. ábra szemlélteti [9].



2.2. ábra: A helyes válaszok aránya csoportonként

A 2.2. ábra alapján megfigyelhető, hogy a generált kérdéssorra kapott helyes válaszok aránya nagyon jól megközelíti a manuálisan készített kérdésekre adott helyes válaszok arányát. Az eredményekből az is látható, hogy az *általános szavak* területéhez kapcsolódó kérdésekre átlagosan jobban tudtak válaszolni a hallgatók.

A következő elemzési lépésben annak a meghatározására törekedtem, hogy van-e kimutatható kapcsolat a szavak osztályozásakor definiált szubjektív koordináták és a helyes válaszok között. A tesztek adatait a 2.1. táblázat tartalmazza [9].

Koordináták\Értékek	a kategória erős ellenpárja	a kategória ellenpárja	a kategória szempontjából irreleváns	a kategória képviselője	a kategória erős képviselője
Relevancia	0 – 0	9 – 26	84 – 124	35 – 79	41 – 84
Specialitás	0 – 0	9 – 26	64 – 124	37 – 79	27 – 84
Értelmesség	0 – 0	0 – 0	28 – 24	92 – 162	53 – 127
Nehézség	0 – 0	5 – 21	41 – 81	125 – 207	2 – 4

2.1. táblázat: A szubjektív koordinátáknak a helyes válaszok számára gyakorolt hatása

A táblázat adataiból jól látható, hogy a szubjektív koordináták közül a *Nehézség* koordináta van legszorosabb kapcsolatban a helyes válaszok számával. Ezen belül is azokra a kérdésekre érkezett a legtöbb helyes válasz, amelyekből kivett szót a szakértő az átlagostól nehezebbnek értékelte.

Az általam végzett kísérletek igazolják, hogy a kifejlesztett kérdésgeneráló mintarendszer alkalmazható a jóval időigényesebb manuális kérdés-előállítás helyett. Az új módszerben alkalmazott többféle hangolási lehetőséggel pedig elérhető, hogy a kérdések pontosan kifejezzék a tesztelés alapjául szolgáló tananyag lényegét.

4. tézis:
Kidolgoztam egy olyan algoritmust, amely az előálló objektív és szubjektív koordináták alapján meghatározza a kérdésként kiemelt szót és annak válaszalternatíváit. Az automatizált kérdésgenerálási módszer gyakorlati alkalmazhatóságának igazolására egy Java nyelven írt szoftvert készítettem. Az elemzések alapján megállapítható, hogy az automatizáltan előállított tesztek eredményei szoros korrelációban állnak a manuális tesztek eredményeivel [9], [14].

3. TOVÁBBI KUTATÁSI FELADATOK

A nagyméretű dokumentumok klaszterezésének legelterjedtebben alkalmazott módszere a BIRCH algoritmus. Ennek feladatspecifikus adaptációját alkalmaztam a tudományos munkám során kidolgozott mintarendszerben. A BIRCH algoritmus alapú klaszterezéssel végzett tesztek eredményeinek alapos tanulmányozását követően felfedeztem néhány lehetőséget, melyekkel speciális esetekben az algoritmus hatékonysága javítható. Ezek közül a legígéretesebb irányokat az algoritmus főbb paramétereinek szoftveresen támogatott beállítása, illetve a nagy elágazási tényezőjű fában való keresés időigényének csökkentése jelentik. Ezekben a továbbfejlesztési témákban elért eredményeimet az értekezés *A.6* számú Függelék tartalmazza.

A szakirodalomban mindmáig nem ismert egzakt módszer a többrétegű neurális hálózatok belső rétegeiben lévő neuronok számának meghatározása. Az alkalmazott neuronok száma a legtöbb esetben konkrét példák alapján való tesztelésekkel kerül meghatározásra. Az optimális számtól kevesebb belső neuron alkalmazása esetén a hálózat nem képes a bemenetek és kimenetek közötti kapcsolatot leíró átviteli függvény reprezentálására. Az optimális számtól több belső neuron alkalmazásával viszont a hálózat a szabályrendszer feltárása helyett hajlamosabb a konkrét példák betanulására. Kutatási munkám

során az alkalmazott háromrétegű neurális hálózat belső rétegében lévő neuronok számát azonosnak választottam a kimeneti rétegben lévő neuronokéval. A munkám speciális alkalmazási területe alapján nyerhető információk alapján lehetőség nyílik a belső rétegben lévő neuronok optimumközeli számának meghatározására. Ezt az osztályozandó dokumentum által képviselt tudományterületnek, a dokumentum méretének, valamint a szavak objektív koordinátáinak alapján végzett tesztekkel kívánom meghatározni.

Annak érdekében, hogy az osztályozásra alkalmazott neurális hálózat tanítását ne kelljen minden dokumentum esetén megismételni, megvalósítottam a hálózat betanult átviteli függvényének igény szerinti mentési és visszatöltési lehetőségét. Mivel azonban a neurális hálózat dinamikusan állítja be a bemeneti rétegében lévő neuronok számát a dokumentum klaszterezésével nyert információk alapján, ezért az elmentett információk csak ugyanolyan stratégiával és paraméterekkel klaszterezett dokumentumok esetén kompatibilisek egymással. A kidolgozott mintarendszer gyakorlati alkalmazhatóságának területe bővíthető lenne egy általános mentési és visszatöltési algoritmus kidolgozásával.

A mintarendszer kifejlesztésével az volt a célom, hogy a feladatsorok idő és emberi erőforrásigényes manuális előállításának folyamatát automatizált módon valósítsam meg. Ehhez olyan kérdések és válaszalternatívák előállítására alkalmas algoritmusokat dolgoztam ki, melyekkel minél pontosabban mérhető a feladatlapot kitöltő személynek az adott témára vonatkozó tudásszintje. A gyakorlati tesztek alapján kimutattam a válaszalternatívaként használt szavak szubjektív koordinátáinak a válasz helyességére gyakorolt hatását a tesztet kitöltő személyek előre ismert tudásszintjének függvényében. Ezáltal olyan információhoz jutottam, mely megmutatja, hogy a kérdezett témát különböző szinten ismerő személyek milyen szubjektív koordinátákkal rendelkező szavakat részesítenek előnyben a helyes válaszok kiválasztása során. Ezeket az információkat is használva lehetőség nyílik a már kidolgozott kérdés, illetve válaszalternatívákat előállító algoritmusok továbbfejlesztésére melyekkel pontosabban mérhető a tesztben résztvevő személyek tudásszintje.

4. SUMMARY

Automated question generation based on an annotated text

The primary motivation of the research constitutes the design and creation of a system model with the application of which predetermined sentence patterns of annotated Hungarian text-documents can automatically be generated. According to the above, in the course of the present research I fulfilled the following tasks.

1. To generate multiple choice questions automatically, I have designed a system model in which a harmonized functioning of a database of a multiple subsystem supports the solution of the task. Output and input interfaces were defined to the modules assigned to fulfil different tasks, which granted the separate check and independent development of the modules (see Chapter 4).
2. Through the clustering of the different concepts represented by the words in the document the criteria describing the similarity of the words were defined. The clustering was carried out both on the basis of the common frequency of the words and their geometric distances measured in the multidimensional space defined by the subjective coordinates of the words. The validity of the model was proved by the width-first, depth-first, best-first search, as well as the task specific implementations of the BIRCH strategies (see Chapter 5).
3. A classification algorithm based on a neural network was designed which is capable of exploring the system of rules describing the subjective and objective connections between the coordinates of the words in the document. With the automated definition of the subjective coordinates of the words the information necessary to highlight the word answering the question and to determine the possible answers was produced (see Chapter 6).
4. On the basis of the information produced task-specific algorithms were developed to generate the answer options. The algorithms fulfil the task applying their own implemented concept hierarchy, subdivision of space and interactive extraction of professional opinions (see Chapter 7).
5. During the automated production of the questions the sentences of the protocol was determined on the basis of quantitative and qualitative criteria of annotations. The practical application of the system model was

tested on humans with different educational background. The results proved that the test papers generated by the automated pattern system measured the intellectual capacities of the tested humans with little or no difference compared to the ones compiled manually. With the above it was proved that the pattern-question generating system developed in the course of my Ph.D. can efficiently substitute the time-consuming manual question generation (see Chapter 8).

New scientific results can be summarized as follows.

Thesis 1: [3], [14]

The elements of the data structure and those of the function model were specified. I defined the system components necessary to work out the automatic question generation, as well as the input and output data related to the system components. The scope of application of the method was specified by determining the constraint conditions. A system-plan to serve the harmonized functioning of the subtasks and a documentary structural model were developed. The potential functioning of the model was demonstrated.

Thesis 2: [1], [7], [8]

The possibilities of clustering methods indicating the distances between the words were analysed and a suitable attribute-system was developed. A Quality Treshold (QT) method based on parameter-driven software was implemented and optimized. The assigned clusters well represent the semantic fields of the words in the Automatic Question Generation (AQG) system of the domain investigated.

Thesis 3: [3], [14]

The objective input coordinates necessary to select the assigned words in the question were defined. A classification algorithm based on a neural network to describe the connection between the objective and subjective coordinates of the words in the document was developed. The precision expected in the Automatic Question Generation (AQG) is achieved by the Counter-Propagation Network (CPN).

Thesis 4 [9], [14]

An algorithm to generate the answer alternatives on the basis of the objective and subjective coordinates of the words assigned as questions was developed. To prove the practical applicability of the automated question generating method, Java-language software was developed. On the basis of the analyses it can be concluded that the results gained from the automated tests closely correlate to the ones of the manual tests.

Saját publikációk az értekezés témakörében

Idegen nyelvű folyóiratban közölt publikációk

[1] Bednarik, L. & Kovacs, L. (2012). Efficiency Analysis of Quality Threshold Clustering Algorithms, Production Systems and Information Engineering, University of Miskolc, Volume 6 (2012), ISSN 1785 – 1270, pp. 15-26.

[2] Kovacs, L. & Bednarik, L. (2011). Methodological Background of Course Material Ontology Models, NEWSLETTER, Precarpathian University PEDAGOGICS, Issue XXXVIII, Beregszász, Ukrajna, pp. 159-164. Вестник Прикарпатского Университета 2011, Ивано-Франковск, Сборник научных трудов "Педагогика, Выпуск XXXVIII" Удк 11.1:37.013.2 страницы: 159-164.

Magyar nyelvű folyóiratban közölt publikációk

[3] Kovács, L. és Bednarik, L. (2012). Osztályozási feladatok a kérdésgenerálási mintarendszerben, A Gépipari Tudományos Egyesület Műszaki Folyóirata (GÉP), ISSN 0016 – 8572, LXIII. évfolyam, 2012/5, pp. 83-86.

[4] Bednarik, L. (2010**).** Digitális oktatási anyagok készítése és megjelenítése, Képzés és Gyakorlat Neveléstudományi szakfolyóirat, Célok és módszerek a tudásalapú társadalom nevelési intézményeiben**,** ISSN 1789 – 8587, Kaposvár, pp. 55-70.

[5] Bednarik, L. (2009). Mi a szövegbányászat?, 50 éves a sárospataki felsőfokú pedagógusképzés, ISSN 0230 – 0435, SPF 26, Sárospatak, pp. 131-135.

Idegen nyelvű konferencia kiadványban közölt publikációk

[6] Kovacs, L. & Bednarik, L. (2010), DITA Metadata and XSLT-FO in Document Management, XXIV**.** microCAD International Scientific Conference**,** ISBN 978-963-661-919-0**,** Miskolc, pp. 79-82.

[7] Kovacs, L. & Bednarik, L. (2011). Extension of HAC clustering method with quality threshold, 9th IEEE International Symposium on Intelligent Systems and Informatics (SISY 2011), Subotica, Serbia, 2011., 10.1109/SISY.2011.6034333, pp. 257 – 261.

[8] Kovacs, L. & Bednarik, L. (2011). Parameter Optimization for BIRCH Pre-Clustering Algorithm, 12th IEEE International Symposium on Computational Intelligence and Informatics (CINTI 2011), Budapest, Hungary, 10.1109/CINTI.2011.6108553, pp. 475 – 480.

[9] Kovacs, L. & Bednarik, L. (2012). Automated EA-type Question Generation from Annotated Texts, IEEE 7th International Symposium on Applied Computational Intelligence and Informatics (SACI 2012), Timisoara, Romania, 10.1109/SACI.2012.6250000, pp. 191 – 195.

Magyar nyelvű konferencia kiadványban megjelent publikációk

[10] Bednarik, L. (2010). Ontológiai technológiák és alkalmazásai, I. KÁRPÁT-MEDENCEI NEMZETKÖZI MÓDSZERTANI KONFERENCIA, Lehetőségek és alternatívák a Kárpát-medencében, ISBN 978-973-9541-13-9, Kaposvár, pp. 154-163.

[11] Bednarik, L. (2010**).** Dokumentumok meta-adatai és a DITA XML dokumentum-formátum**,** „Szellemi tőke, mint versenyelőny" konferencia**,** lektorált CD kiadvány, ISBN 978-963-216-270-6, Komárom, Szlovákia, pp. 212-227.

[12] Kovács, L. és Bednarik L. (2010), Digitális dokumentumok formátumai és az XSLT-FO, Multimédia az oktatásban konferencia, MTESZ Neumann János Számítógép-tudományi Társaság, In: Berke József (szerk), CD kiadvány, ISBN: 978 615 5036 04 0, Nyíregyháza, 2010.

[13] Bednarik, L. és Kovács, L. (2011). Tananyag-ontológia modell és a DITA keretrendszer, IV. MISKOLCI TANI-TANI KONFERENCIA, Pedagógia és társadalmi felelősségvállalás, ME BTK TANÁRKÉPZŐ INTÉZET, 2011.
http://www.tanarkepzo.hu/sites/default/files/mtk_iv_program.pdf

[14] Bednarik L. (2012). Automatizált kérdésgeneráló mintarendszer, Informatikai konferencia, ME Comenius Főiskolai Kar, Real Tudományok Intézete, Informatika, ME CFK Könyvtára, Konferencia-előadás kézirata, pp. 1-10.
http://www.ctif.hu/public/bednarikpublic.htm

[15] Bednarik, L. (2006). Az e-learning fázisainak Web-es megvalósítása az oktatásban, MAGYAR TUDOMÁNY NAPJA 2006 konferencia, ME Comenius Tanítóképző Főiskolai Kar, ME CFK Könyvtára, Konferencia-előadás kézirata, pp. 1-15.
http://www.ctif.hu/public/bednarikpublic.htm

Lektorált hazai kiadvány

[16] Bednarik, L. (2010). Oktatás- és információtechnika I., Miskolci Egyetem Comenius Tanítóképző Főiskolai Kar, ME CFK Könyvtára, Magyar Tudományos Művek Tára, pp. 1-56.

[17] Bednarik, L. (2004). Elektronikus tanulást segítő WEB-es mintarendszer kidolgozása, Miskolci Egyetem, Gépészmérnöki és Informatikai Kar, Általános Informatika Tanszék, Diplomamunka, Konzulens: Kovács., L., pp. 1-98.

Közlésre benyújtott, értékelés alatt álló publikációk

Kovacs, L., Gyöngyösi, E. & Bednarik, L. (2012). Development of classification module for automated question generation framework, Teaching Mathematics and Computer Science, Debrecen, Hungary, ISSN 1589 – 7389, *submitted.*

Kovacs, L. & Bednarik, L. (2012). Application of Semantic Clustering in Question Generation Engine, Transactions on Automatic Control and Computer Science, Scientific Bulletin of the "Politehnica" University of Timisoara, Romania, ISSN 1224-600X, *submitted*.

Hivatkozások

[Brown, 2004] Jonathan Brown, Maxine Eskenazi (2004). Retrieval of Authentic Documents for Reader-Specific Lexical Practice, In: Proceedings of InSTIL/ICALL Symposium 2004. Venice, Italy.

[Collins, 2007] Collins, M., Duffy, N. (2007). New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron, Proc. of 40th Annual Meeting of the Association for Computational Linguistics, pp. 263-270.

[Coniam, 1997] Coniam, D. (1997). A Preliminary Inquiry into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests, CALICO Journal, 14 (2-4), pp. 15-33.

[Fajszi, 2004] Fajszi, B., Cser, L. (2006). Üzleti tudás az adatok mélyen, IQSYS Informatikai Rt, Budapest, 2004, pp. 179-181.

Futó, 1999] Futó, I. (1999). Mesterséges Intelligencia. Aula Kiadó, Budapest, 1999.

[GermaNet, 2009] GermaNet (2009). GermaNet: Introduction, Eberhard Karls Universität Tübingen, http://www.sfs.uni-tuebingen.de/GermaNet/.

[Gütl, 2008] Gütl, C. (2008). Automatic Limited-Choice and Completion Test Creation, Assessment and Feedback in modern Learning Processes, Paper read at LRN Conference 2008, Guatemala, February 12th – 16th.

[Gütl, 2011] Gütl, C., Lankmayr, K., Weinhofer, J & Höfler, M. (2011). Enhanced Automatic Question Creator – EAQC: Concept, Development and Evaluation of an Automatic Test Item Creation Tool to Foster Modern e- Education, The Electronic Journal of e-Learning, ISSN 1479-4403, Volume 9, Issue 1 2011, pp. 23-38.

[Miller, 1990] Miller, G., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. (1990). Five Papers on WordNet. CSL Report 43. Cognitive Science Laboratory. Princeton University.

[Piwek, 2008] Piwek, P., Prendinger, H., Hernault, H., & Ishuzuka, M. (2008). Generating questions: An inclusive characterization and a dialogue-based application. Workshop on the Question Generation Shared Task and Evaluation Challenge. NSF, Arlington.

[Piwek, 2009] Piwek, P., Mostow, J., Chali, Y., Forascu, C. & Gates, D. (2009). The Question Generation Shared Task and Evaluation Challenge. In Workshop on the Question Generation Shared Task and Evaluation Challenge, Final Report, The University of Memphis: National Science Foundation.

[Rus, 2007] Rus, V., Cai, Z. & Graesser, A. C. (2007). Experiments on Generating Questions About Facts, Computational Linguistics and Intelligent Text Processing, 8th International Conference, CICLing 2007, Mexiko City, Mexiko, February 18-24, 2007, Proceedings, pp. 444–455.

[Sumita, 2004] 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.

[WordNet, 2010] WordNet (2010). WordNet: A lexical database for English, Princeton University, http://wordnet.princeton.edu.