A brit Nemzeti Szövegbányászati Központ

A *National Centre for Text Mining (NaCTeM) – Nemzeti Szövegbányászati Központ* a Manchesteri, a Liverpooli és a Salfoldi Egyetemek konzorciuma, amelynek több önfinanszírozó partnere is van az Egyesült Államokból és Japánból. A NaCTeM tevékenységét az élettudományok, ezen belül az orvosi, biológiai terület szövegeire összpontosítja, mivel növekedni látja a biológiai témájú szövegek bányászata és azok keresésével, hozzáférésével, feltárásával és integrálásával kapcsolatos automatikus eljárások iránti igényt. A konzorciumot 2004-ben alapították. Az anyagi támogató Joint Information Systems Committee (JISC), a Biotechnology and Biological Sciences Research Council (BBSRC) és az Engineering and Physical Sciences Research Council (EPSRC) által adott összeghez a konzorcium is ugyanannyit tesz hozzá.

Az élettudományi, az orvosi és a biológiai szakterület dinamikus fejlődése mennyiségben és tematikában egyaránt hatalmas mértékben növekvő irodalmat állít elő. Eredményes automatikus feldolgozásra van igény, ami segít meghatározni, összegyűjteni és használni az elektronikus formában elérhető szakirodalmat. Noha jelentős adatokat tárolnak orvosi, biológiai tényadatbázisokban, a legrelevánsabb és leghasznosabb információ az adott szakterületek szakirodalmában található. A Medline 14 millió rekordot tartalmaz, és állománya havi negyvenezer referátummal gyarapodik. Vannak nyílt hozzáférést nyújtó kiadók is (mint például a BioMed Central), amelyek növekvő teljes szövegű cikkadatbázist birtokolnak. Mindemellett növekszik a biológiai tényadatbázisok és a szakirodalom összekapcsolásának igénye, és az erre vonatkozó tevékenység is, melynek célja az adatbázisok kiegészítése és a szakirodalom igazolása. Ez ma meglehetősen munkaigényes tevékenység; többnyire manuálisan, néhány bonyolult eszköz használatával végzik, és nem lebecsülendő a hibázás vagy fontos elemek kihagyásának veszélye sem. Szintén növekvő igény mutatkozik a biológusok között arra, hogy a tényadatbázisokat a szakirodalommal és saját előállított adataikkal együttesen tárják fel. Így a szakirodalmi szövegbányászat nem választás kérdése, hanem a hatékony tudáskinyerés, -menedzsment, fenntartás és frissítés eszköze. A biológusok a Medline kereséséhez hagyományosan a PubMed kézzel, kötött tezaurusz segítségével indexelt adatbázisát használják, Boole-algebrai kifejezések segítségével. Gondot jelent, hogy az ebben használt hagyományos információkereső eljárások nem számolnak a szinonimákkal és az azonos alakú szavakkal. Az ellenőrzött indexelés problémája, hogy nem követi a dokumentumok dinamikáját.

A hagyományos visszakereséssel túl nagy a találati halmaz, ezért használata nehézkes. Az irodalomban megjelenő nagyszámú új szakkifejezés miatt a szövegbányászati eszközök, pl. az automatikus kifejezésmenedzselő eszközök nélkülözhetetlenek a módszeres és hatékony orvosi, biológiai adatok indexelésen és visszakeresésen túlmutató gyűjtéséhez. A kézzel szabályozott szótárak hibáznak, szubjektívek, és a témalefedésben korlátozottak.

Az orvosi, biológiai irodalomkutatás számos technikai és nyelvészeti kihívást jelent. Technikait pl. a hozzáférés körülményessége (a sok különféle, és nem szöveges formátum miatt: táblázatok, képek, ábrák), nyelvit pedig az orvostudomány és biológia részterületeinek különböző nyelvhasználata miatt. Így az egyik legnagyobb kihívás a biológiai terminológia. A cikkek megértéséhez először a bennük lévő szakkifejezések pontos definíciójára van szükség. Új kifejezések viszont naponta kerülnek be az egyes szakterületek kifejezésszótáraiba, ráadásul nem is biztos, hogy a világ különböző laboratóriumaiban pontosan ugyanazokat a terminusokat használják. Így e szótár kézi karbantartása lehetetlen. Legalább háromszáz orvostudományi adatbázis kínál terminológiai tartalmat, ezek közül sok deszkriptorokat használ a cikkben szereplő kifejezések helyett, és ez megnehezíti a tartalomelemzést. A terminológiai feldolgozás (azaz a kifejezések azonosítása, osztályozása és kapcso-

latainak vizsgálata) jelenti az orvostudományi szövegbányászat szűk keresztmetszetét, ami jelentősen csökkenti a „magasabb szintű” szövegbányászat lehetőségeit. Több megközelítés létezik a folyó szövegekben való automatikus kifejezésfelismerésre. (Fontos, hogy megkülönböztessük az itt használt technikai kifejezéseket a dokumentumok jellemzésére használt indexkifejezésektől. A jó indexkifejezés nem feltétlenül technikai kifejezés. A technikai kifejezés áll ugyanis a szövegelemzés érdeklődésének középpontjában, még ha az nem is gyakran jelenik meg a kifejezésgyűjteményben. A dokumentumban lévő minden technikai kifejezés érdekes lehet a szövegelemzés szempontjából.)

A szövegben szereplő kifejezések felismerése nem a végső cél. Ezeknek illeniük kell a meglévő tudáshoz és/vagy egymáshoz is. Osztályozni és hierarchiába kell rendezni őket, mivel ezek teremtik meg a szavak világa, valamint az ontológiák és más osztályozási sémák világa közötti kapcsolatot. Az ontológiai elemek határozzák meg azt a térképet, amely lehetővé teszi az ontológián alapuló információkinyerést.

A kifejezések közötti kapcsolatok kinyerésére a leggyakoribb módszer a részleges mondatelemzés és az információkinyerés (Information Extraction = IE). Ezek részben a mintamegfeleltetésen, részben az IE-alapú szemantikai sablonokon alapulnak. A mintamegfeleltetés megközelítés jellemzően eredményes, de a doménorientált minták létrehozása meglehetősen költséges. A visszahívást befolyásolhatja, ha a mintának nem elég széles a lefedése. Mivel a statisztikai, tudásintenzív vagy gépi tanulási megközelítések külön-külön használata nem felel meg a használók szemantikai jellemzők iránti igényeinek, ezek együttes használata lehet célravezető. Az alkalmazott módszernek egyben dinamikusnak is kell lennie, hogy kövesse a terminológia gyors változásait. A legtöbb jelenlegi rendszer az ismert kapcsolatokat használja, és a szemantikus vagy fogalmi entitásokat, entitáselemeket, entitástulajdonságokat alkalmazó tényinformációkat célozza. A konzorcium nemcsak az entitások, tulajdonságok és tények leszűrését, hanem az adatbányászat során a társulások és a kapcsolatok felfedését is meg fogja valósítani.

A Nemzeti Szövegbányászati Központ (NaCTeM) szerepe

A NaCTeM legfőbb feladata, hogy kiváló szolgáltatásokat hozzon létre és nyújtson az Egyesült Királyság kutatási szférájának, különös tekintettel a biológiai és orvostudományi területen. A megfelelő eszközök értékelése és kiválasztása folyamatban van, szem előtt tartva a partnerek és megrendelők igényeit, figyelve a versenytársakra és a technológiaszolgáltatókkal való kapcsolatban rejlő előnyökre.

A NaCTeM főbb céljai:

1. Teljes körű tanácsadás, szolgáltatás, információtovábbítás, oktatás, adatszolgáltatás a friss technológiákról, konzultációs szolgáltatásokról, tananyagokról, demonstrátori projektekről, elsősorban a brit szövegbányászok számára.
2. Nemzeti és nemzetközi esemény-nyilvántartás a szövegbányászat területén.
3. A biológiai vonatkozású szövegbányászat legjobb gyakorlatainak konszolidálása, kifejtése és közzététele más szakterületek számára.
4. A szövegbányászat széles körű tudatosítása, és részvétel biztosítása minden tudományos diszciplína számára, ideértve az üzlet és a menedzsment területét is.
5. Kapcsolattartás és -fejlesztés az alkalmazások és eszközök szolgáltatóival a legjobb gyakorlat és ellátás megteremtéséért.

A NaCTeM 2005 nyarára egy interdiszciplináris biológia-központot kíván létrehozni, ahol élettudományi kutatók, fizikusok, kémikusok, matematikusok, számítástudományi szakemberek, számítógépes szakemberek és nyelvmérnökök működhetnek együtt az ágazat szolgáltatóival és eszközfejlesztőivel.

A konzorcium portálja tartalmazni fogja a szolgáltatáskínálatot. Ennek ágai: eszközökhöz, forrásokhoz és támogatáshoz való hozzáférés segítése, a források és eszközök online hozzáférése és teljes alkalmazáscsomagok. A szolgáltatások tartalma a következő lesz:

- hozzáférés az élvonalbeli fejlesztők legfrissebb szövegbányászati eszközeihez és a forgalmazott termékekhez;
- hozzáférés ontológiai könyvtárakhoz;
- hozzáférés nagy és változatos kínálatú adatforrásokhoz (tanácsadás és beszerzés több termékkategóriában);
- hozzáférés az adatszűrő eszközök könyvtárához;
- online tananyagok, közlemények és szakanyagok;
- szövegbányászati eszközökkel kapcsolatos online tanácsadás speciális igények esetén;

- a szövegbányászat és reprezentálás GRID-alapú rugalmas eszközeinek a használó általi online bemutatása; együttes eszköz-, forrás- és adatkínálat a portálon való szövegbányászat céljából;
- terjesztő és marketingtevékenységek, pl. oktató- és tananyagok, konferencia- és workshopszervezés;
- szövegbányászati eszközök, annotált szövegkorpuszok és ontológiák közös fejlesztése, kiterjesztése;
- szövegbányászati eszközök kipróbálása és értékelése.

A konzorcium eleinte a kutatási szféra munkatársaira, majd a biotechnológiai és az orvosi, biológiai szolgáltatási lánc üzleti vállalkozásaira számít használókként. Ezenfelül érdeklődés várható a közszféra információs szervezeteitől, a bioipar kis- és középvállalataitól és az IT-szektorból, regionális fejlesztési ügynökségektől, egészségügyi szolgáltatóktól és hivataloktól, gyógyszergyáraktól, valamint olyan további területek képviselőitől, mint az élelmiszeripar, a kormányzat és a média.

A NaCTeM által átfogott területek: bioinformatika és genomika (alapvető cél a kísérleti adatok, genominformáció és biomedikai szakirodalom szimultán használói értelmezését segítő stratégia és módszer felfedezése), ontológiák, szókészlet és annotált szövegkorpuszok.

Az ontológiák a szakterület-specifikus tudást ábrázolják, és segítik az információcserét. Az információt strukturált formában tárolják (egyben a szemantikus web alapvető forrásai). Egy olyan növekvő szakterületen, mint amilyen az orvostudomány és a biológia is, nagyon fontos az ontológia bővíthetősége. Mivel az ontológiák az automatikus orvosi, biológiai tudásszerzés céljából szükségesek, a kihívást automatikus frissítésük jelenti. Megoldást a kifejezésklaszterezés és -osztályozás jelenthet. A kifejezésklaszterezés a kapcsolatban álló kifejezések közötti lehetséges összefüggéseket azonosítja, amelyek felhasználhatók az ontológiákban a szemantikus relációpéldák frissítésére vagy igazolására. A kifejezésosztályozás eredménye egy ontológia taxonómiai szempontjainak igazolására vagy frissítésére használható.

A lexikai források (szótárak, szószedetek, taxonómiák) és annotált szövegkorpuszok szintén fontos elemei a szövegbányászatnak. Az elektronikus szótárak formális nyelvészeti információt adnak a szavak formájáról. Mivel az ontológiák fogalmakat jelenítenek meg, és nincsen kapcsolatuk a szavak felszíni világával, összekötő eszköz szükséges a kanonizált szöveges karakterláncok (szavak) és az ontológiai fogalmak közé. Szótárak és taxonómiák segítenek a térképezésben. Annotált szövegkorpuszok (GENIA) a szabálykészítéshez, a gépi tanulási módszerek oktatásához és az értékeléshez kellenek.

A használói igények szolgálata

Feltételezhető, hogy több használó szeretné többször ugyanazon adatbázisokat – pl. a Medline-t – részben ugyanolyan elemzési módszereknek alávetni. S ha már egy feldolgozás megtörtént, akkor újabb esetekben elegendő csak az ügyfél kívánságának megfelelő kisebb profilmódosítás szerint elemezni. Az elemzés – főként az ügyfél által kívánt gyors módja – nagy gépi kapacitást köt le, ezért ezt GRID-alapon lehet majd a portálon keresztül igénybe venni. Ha a portálon a használók legalábbis részben hasonló feladatokat végezhetnek, és szolgáltatásokat kapnak, fontossá válik a biztonságos és bizalmas hozzáférés is.

A szolgáltatásfejlesztés két hangsúlyos pontja a skálázhatóság és a hatékonyság. A használók ugyanis egyéni igényeiknek megfelelő szolgáltatásokat kívánnak, pl. megállhatnak a visszakeresés és ténykinyerés után, vagy csupán mondatszétválasztást, kijelöléseket, dokumentumfelosztást kívánnak. A munkaelemenkénti szolgáltatás a szabványosítás kérdését veti fel, nemcsak a transzportprotokoll szintjén, hanem nyelvészeti szinten is, pl. nyelvi adatelemek és attribútumok következetes címkézése és interpretálása érdekében.

A használók várhatóan részben bioinformatikai szakértők, részben pedig adott szakterületek információtudományban járatlan szakértői lesznek. Így a használók különféle csoportjai számára meglévő tudásukhoz igazodó eligazító anyagokat is szükséges készíteni. A konzorcium az ügyfélszolgálat-értékelés minősége terén a konzorciumi tag Genfi Egyetem kiterjedt tapasztalataira támaszkodhat.

Sok szövegbányászati terméknek nem elég fejlett a kifejezésmenedzsmentje. A konzorcium az ATRACT (Automatic Term Recognition and Clustering for Terms = automatikus kifejezésfelismerő és klaszterező rendszer) adaptálásában látja a megoldást, amelyet a konzorciumi tag Salfold Egyetem fejleszt. A rendszer kipróbált, nyelvtől független hibrid statisztikai és nyelvészeti

alkalmazás, amely integrálható a szövegbányászati folyamatba.

Az ontológiaalapú információkinyerés egyelőre gyerekkorát éli. A NaCTeM által fejlesztett ontológia a Manchesteri Egyetem CAFETIERE (Conceptual Annotations for Facts, Events, Terms, Individual Entities, and RElations = események, szakkifejezések, egyedi entitások és kapcsolatok fogalmi annotációja) információkinyerő termékén alapul. A rendszer skálázható. Szabályalapú elemző, szabályai kapcsolhatók a használók ontológiáihoz és eseményeihez. Szabályalkotásának figyelemre méltó teljesítménye csökken ennél a megközelítésnél, mivel a szabályíró kevesebb és sokkal általánosabb szabályokat ír, a mások által írt ontológiáknak köszönhetően. Így közvetlen, kedvező kapcsolat van az ontológia építőjének világa és az információt leszűrő világa között. A CAFETIERE képes továbbá időben ismétlődő keresésekre is, ami nemcsak az üzleti hírszerzés célú efemer kishírek kutatásának szempontjából érdekes, hanem bármely gyorsan változó terminológiájú és tudású szakterület kutatására is hasznos lehet. A tapasztalat szerint ugyanis van igény csak adott (múltbéli) időpontokban releváns tudás kinyerésére is.

A Liverpooli Egyetem a Berkeleyvel közösen kifejlesztette a Cheshire harmadik generációs, nemzetközi szabványokon alapuló online információkereső rendszert, amelyet számos nemzeti szolgáltatás és projekt használ az Egyesült Királyságban. A Cheshire-t a NaCTeM adatok aratására és indexelésére fogja használni egy fejlett klaszterezési eljáráson belül, amely lehetővé teszi az egységek automatikus keresztlinkelését és gyors visszakeresését. A Cheshire fejlesztési munkái a szövegbányászat és az információ-visszakeresés igényeire összpontosulnak, különös tekintettel a metaadatok fejlett indexelésére és visszakeresésére, az indexkifejezések továbbfejlesztett mérésére, keresési interfészekre és az ontológiamenedzsmentre. A legfontosabb fejlesztés a SKIDL lesz (SDSC Knowledge and Information Discovery Lab = SDSC tudás és információ-visszakereső laboratórium). Ez az eszközrendszer nemcsak a hagyományos információkinyerésre lesz képes, hanem kapcsolatokat épít olyan biológiai entitások között, mint például elemzett genomadatok, biológiai tudományos szövegek vagy bibliográfiai adatok. A Cheshire legfőbb erénye, hogy képes a hibrid szövegbányászatra (például folyóiratból és a DNA szöveges reprezentációjából), átlátható és hatékony módon.

Az adatbányászatot hagyományosan olyan területeken használták, ahol a strukturált adatok voltak jellemzőek, például ügyfélszolgálati menedzsment (CRM), banküzlet és kiskereskedelem. E tevékenységek fókusza a szöveg mélyén lévő ismeretlen, de hasznos tudás felfedezése. A szövegbányászat kiterjeszti szerepét a szöveges dokumentumok félig strukturált és strukturálatlan világára.

A NaCTeM szolgáltatásai nem lesznek azonnal elérhetők. A munkát három évre tervezik, fokozatos fejlesztésekkel, a teljes szolgáltatási kapacitás eléréséig. Kezdetben, a fejlesztési szakaszban clearinghouse-ként fog működni, szövegbányászati eszközök és irodalom webes katalógusaként és tárházaként. Emellett tanácsadást és oktatást is végeznek. Az e munka során begyűjtött visszajelzések és kapcsolatok alapján fejlesztik a későbbi szolgáltatásokat. Folyamatosan keresik a kapcsolatokat a lehetséges használókkal minden egyéb szakterületről is. Eseménykalendáriumuk jelzi eddigi kötődéseiket és tevékenységük alakulását.

/ANANIADOU, Sophia et al.: The National Centre for Text Mining: aims and objectives. = Ariadne, 42. köt. 2005. január, http://www.ariadne.ac.uk/issue42/ananiadou/

(Mikulás Gábor)