Gépi tanulási módszereken alapuló kulcsszókinyerés és alkalmazási lehetőségei eltérő doméneken

Berend Gábor

Konzulensek:
Dr. Csirik János
Dr. Farkas Richárd

Szegedi Tudományegyetem
Informatikai Doktori Iskola

PhD értekezés tézisei

Szeged, 2014

Motiváció

Mindennapi tevékenységeink során egyre változatosabb tartalmú és formájú adatokat hozunk létre, melyben nagy szerepe van a különféle okos eszközöknek. Ezen adatok feldolgozása – keletkezési ütemükből, valamint volumenükből adódóan – csakis gépi módszerekre támaszkodva képzelhető el. Így váltak a szöveges tartalmak feldolgozásának és elemzésének is nélkülözhetetlen eszközeivé a *természetesnyelv-feldolgozás* által kínált különféle automatizált megoldások. Ezek közé a megoldások közé tartozik az ún. *kulcsszókinyerés* is, mely során szöveges dokumentumokhoz azok tartalmát tömören jellemezni képes kifejezések halmazát rendeljük számítógépes segítséggel.

A kulcsszavazási feladat céldokumentumai tetszőlegesek lehetnek: célul tűzhető ki többek között (mikro)blogbejegyzések, újsághírek vagy tudományos publikációk kulcsszavainak meghatározása. A különféle doménbe tartozó dokumentumok természetszerűleg különféle sajátosságokkal is rendelkeznek stílusuknak, strukturáltságuknak megfelelően. Éppen ezért érdemes a kulcsszavazási feladat automatizált megoldása során olyan gépi tanulási modelleket készíteni, melyek a céldokumentumok stílusbeli sajátosságait is figyelembe veszik.

A szerző disszertációjában különböző doménekből származó, magyar és angol nyelven írt szöveges dokumentumok kulcsszavainak gépi tanuláson alapuló, automatizált meghatározására képes modellek létrehozásával foglalkozott, illetőleg az ezen modellek által nyert kimenetek további felhasználhatóságát vizsgálta meg. Míg a korábbi hasonló munkák jellemzően angol nyelvű tudományos publikációk kulcsszavainak automatikus meghatározásával foglalkoztak, addig jelen tézis szerzője angol és magyar nyelvű szövegekkel egyaránt dolgozott, valamint a (hagyományosnak mondható) tudományos publikációkon túlmenően a termékvéleményezések, illetve újsághírek kulcsszavainak meghatározásának problémájára is adott megoldásokat.

A disszertáció felépítése

A disszertáció két fő részre bontható. Az első részben a szerző különféle doméneken és nyelveken végrehajtott kulcsszavazási feladatokra adott megoldásait mutatja be. A dolgozat második logikai egysége az előzőekben bemutatott kulcsszavazási modellek további alkalmazásokba történő integrációjával foglalkozik.

A 3. fejezetben a kulcsszavak újsághírekből történő kinyerésével foglalkozik a szerző. Az újsághírek sajátosságai, hogy tartalmukra és stílusjegyeikre nézve egyaránt nagyfokú heterogenitást mutatnak. Az ebben a fejezetben bemutatott eljárás által feldolgozott dokumentumok heterogenitásukon túl azzal a sajátossággal is rendelkeznek, hogy azok – a többi fejezetben használt dokumentumoktól eltérően – magyar nyelven íródtak.

A 4. fejezet a standard kulcsszavazási feladat (ti. angol tudományos publikációk kulcsszavainak meghatározása) megoldására ad újszerű jellemzőkön alapuló gépi tanulási modellt. Az új jellemzők a kulcsszójelöltekre vonatkozó szemantikus világtudás reprezentálásra hivatottak, melyek meghatározására a szövegek nyelvi elemzése és külső erőforrások fölhasználása útján került sor.

Az 5. fejezetben a korábban kezelt dokumentumtípusoktól jelentősen eltérő dokumentumok, termékvéleményezések véleménykifejezéseinek meghatározására létrehozott modelljét mutatja be a szerző. Ebben a fejezetben alátámasztást nyer a szerző azon feltételezése, miszerint a szubjektív véleménykifejezések kinyerése során a standard (objektív) kulcsszavak kinyerésére kifejlesztett eljárások sikerrel alkalmazhatók. A fejezetben bemutatásra kerülnek továbbá olyan doménspecifikus jellemzők is, melyek számottevő javulást eredményeztek a kinyert véleménykifejezések minőségében.

A 6. fejezetben a korábbiakban bemutatott kulcsszavazási eljárások további alkalmazásokban történő felhasználási lehetőségeire tér ki a szerző. A fejezetben egy dokumentumhalma-

			Fejezet			
Megjelenés	Év		3	4	5	6
SEMEVAL	2010	[3]	•			
RANLP	2011	[9]	•			
NLE	2014	bírálat alatt	•			
ECAI	2010	[8]		•		
IJCNLP	2011	[1]			•	
RANLP	2011	[7]			•	
WASSA	2012	[6]			•	
CICLing	2013	[5]				•
IJCNLP	2013	[4]				•

1. táblázat. A disszertáció fejezetei és a hivatkozott saját publikációk közötti kapcsolat.

zok kulcsszavainak egyidejű meghatározására irányuló eljárás, valamint egy vizualizációs keretrendszer kerül bemutatásra.

Az 1. táblázatban találhatók a szerző disszertációjához kapcsolódó publikációi. A szerző további publikációi megtalálhatók a `http://www.inf.u-szeged.hu/~berendg/?pp=publ` URL-en.

A tézisek eredményeinek összegzése

A disszertáció elsődleges célja a különböző doménekből (újsághírek, tudományos publikációk, termékvéleményezések) származó dokumentumok kulcsszavainak automatikus meghatározására alkalmas algoritmusok bemutatása. Az egyes domének szövegeinek feldolgozására javasolt eljárások figyelembe veszik azok sajátosságait, valamint elmondható róluk, hogy nagyban támaszkodnak a kulcsszavazandó dokumentumon kívülről származó külső információkra. A következőkben rövid bemutatásra kerülnek a tézispontok főbb eredményei.

Újsághírek kulcsszavazása

A 3. fejezetben az Origo hírportál szöveges archívumában található dokumentumokhoz történő kulcsszavak hozzárendelésére irányuló munkáját mutatja be a szerző. Az itt bemutatott kulcsszavazási feladat több sajátossággal is rendelkezett. A kulcsszavazó eljárásnak egyrészt tudnia kellett kezelni a magyar nyelv morfológiai gazdagságából adódó sajátosságokat. Ezen felül külön fontossággal bírt, hogy a kinyert kulcsszavak ne csupán az egyes dokumentumok leírására legyenek alkalmasak, hanem a teljes hírarchívum tekintetében is koherensen viselkedjenek (pl. a szinonim jelentésű kulcsszavak elkerülésével). További jellemzője volt a feladatnak azon ún. *absztrakt kulcsszavak* kezelésének a kiemelt fontossága. Az absztrakt kulcsszavak úgy képesek egy-egy dokumentum tartalmának összefoglalására, hogy abban nem találhatók meg. A fejezetben bemutatott rendszer kiértékelése alapján a meghatározott kulcsszavak a dokumentumok 75.44%-ában érték el a kívánatos minőséget, mely eredmény jelentősen meghaladta az Origo hírportál által támasztott előzetes elvárásokat. Korábbi publikációja [8] alapján a szerző a tézis legfontosabb saját eredményeinek a következőket tekinti:

- A kulcsszójelöltek rangsorolására létrehozott módszer
- Absztrakt kulcsszavak meghatározása a Wikipédiából kinyert definíciók alapján
- Absztrakt kulcsszavak meghatározása a Wikipédia linkstruktúrája alapján

Tudományos publikációk kulcsszavazása

A 4. fejezetben újszerű szövegen kívüli jellemzőkre támaszkodó kulcsszavazó modelljét mutatta be a szerző. A bevezetett jellemzők nem csupán a tudományos publikációk doménjén voltak alkalmazhatók, hanem – ahogy azt az 5. fejezetbeli alkalmazásuk mutatta – doménfüggetlen tulajdonsággal bírtak, ami

széleskörű alkalmazhatóságukat vetíti előre. A javasolt modell hasonlóan vagy jobban teljesített 2 standard benchmark tudományos publikációkat tartalmazó adatbázison is a jelenleg elérhető kulcsszavazó rendszerekhez képest.

A szerző kulcsszavazási megoldásában dokumentumon kívüli információra is támaszkodott, melynek fő forrása a Wikipédia volt. Korábbi munkákban ugyan találkozhattunk már hasonló elképzelésekkel, fontos különbség azonban, hogy míg a szerző kizárólag a Wikipédia kategóriastruktúráját fölhasználva épített be külső tudást rendszerébe, addig más munkák a Wikipédia összes szöveges tartalmának feldolgozását és indexelését tették szükségessé, jóval erőforrásigényesebbé téve ezáltal a hasonló megközelítéseket. Korábbi publikációi [2, 3, 9] alapján a szerző a tézis legfontosabb saját eredményeinek a következőket tekinti:

- A korábbi kulcsszójelölt-állítási stratégiák kiterjesztése
- Szekvenciákat kódoló jellemzők alternatív reprezentációjának bevezetése
- Szövegen kívüli információk kiaknázása

Véleménykifejezések kinyerése

Az 5. fejezetben igazolást nyert a szerző azon hipotézise, mely szerint a kulcsszókinyerési technikák adaptálhatók a véleménykifejezések kinyerésére irányuló feladat megoldása során. Mindezek mellett a szerző a kinyert véleménykifejezések minőségét jelentősen javítani képes doménspecifikus jellemzőket is bemutatott tézisében. A különböző terméktípusok véleménykifejezéseinek kinyerésének átjárhatóságát is vizsgálta a szerző, mely során doménadaptációs kísérleteket hajtott végre. Korábbi publikációi [1, 7] alapján a szerző a tézis legfontosabb saját eredményeinek a következőket tekinti:

- Standard kulcsszavazási feladat kiterjesztése véleménykifejezések kinyerésére

- Termékvéleményezésekből és a hozzájuk tartozó véleménykifejezésekből álló korpusz létrehozása
- Doménadaptációs vizsgálatok a különböző termékcsoportok véleménykifejezéseinek kinyerése közötti átjárhatóság biztosítására

Kulcsszókinyerés alkalmazásai

A 6. fejezetben kulcsszókinyerő rendszerének végalkalmazásait ismertette a szerző. A fejezetben bemutatásra került egy információelméleti alapokon nyugvó módszer dokumentumcsoportok kulcsszavainak meghatározására. A fejezet rámutatott arra is, hogy miként lehet korpuszon belüli témákat detektálni a dokumentumok kulcsszavai közötti átfedés vizsgálata útján, illetve hogy miként lehet ezt az információt korpuszvizualizáció során fölhasználni. Korábbi publikációi [4, 5] alapján a szerző a tézis legfontosabb saját eredményeinek a következőket tekinti:

- Dokumentumhalmazok kulcsszavainak meghatározására irányuló eljárás
- Dokumentumok kulcsszóalapú reprezentációjának bevezetése
- Kulcsszóhasonlósági gráf alapján történő klaszterezés és korpuszvizualizáció

Irodalomjegyzék

[1] Gábor Berend. Opinion expression mining by exploiting keyphrase extraction. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1162–1170, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing. URL `http://www.aclweb.org/anthology/I11-1130`.

[2] Gábor Berend. Exploiting extra-textual information in keyphrase extraction. *Natural Language Engineering*, page to appear, 2014.

[3] Gábor Berend and Richárd Farkas. SZTERGAK: Feature engineering for keyphrase extraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 186–189, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/S10-1040`.

[4] Gábor Berend and Richárd Farkas. Keyphrase-driven document visualization tool. In *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, pages 17–20, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. URL `http://www.aclweb.org/anthology/I13-2005`.

[5] Gábor Berend and Richárd Farkas. Single-document

keyphrase extraction for multi-document keyphrase extraction. *Computación y Sistemas*, 17(2):179–186, 2013.

[6] Gábor Berend and Veronika Vincze. How to evaluate opinionated keyphrase extraction? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 99–103. Association for Computational Linguistics, 2012.

[7] Gábor Berend, István T. Nagy, György Móra, and Veronika Vincze. Inter-domain opinion phrase extraction based on feature augmentation. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 41–47, Hissar, Bulgaria, September 2011. RANLP 2011 Organising Committee. URL `http://aclweb.org/anthology/R11-2006`.

[8] Richárd Farkas, Gábor Berend, István Hegedűs, András Kárpáti, and Balázs Krich. Automatic free-text-tagging of online news archives. In *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pages 529–534, Amsterdam, The Netherlands, The Netherlands, 2010. IOS Press. ISBN 978-1-60750-605-8. URL `http://dl.acm.org/citation.cfm?id=1860967.1861071`.

[9] István Nagy T., Gábor Berend, and Veronika Vincze. Noun compound and named entity recognition and their usability in keyphrase extraction. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 162–169, Hissar, Bulgaria, September 2011. RANLP 2011 Organising Committee. URL `http://aclweb.org/anthology/R11-1023`.