Terminológiakivonatolás francia nyelvű szabadalmak leírásaiból különböző módszerek segítségével

Nagy Ágoston[1,2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
nagyagoston@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola,
Francia nyelvészet alprogram
nagyagoston@lit.u-szeged.hu

Kivonat: A cikk egy francia nyelvre készült saját, elsősorban szabályalapú, de statisztikai szűrőkkel is rendelkező terminológiakivonatoló leírását és eredményeit tartalmazza. Célunk annak feltárása, hogy a tisztán szabályalapú terminológiakivonatoló főmodulon kívül alkalmazott szabályalapú és statisztikai módszerek milyen mértékben járultak hozzá a fedés és a pontosság növeléséhez (vagy csökkenéséhez). A terminusok szabályalapú kinyerése véges állapotú automatával történik, a kimenet szűrése pedig először *stopword*-listával, majd tulajdonnév-felismerő modul alkalmazásával. A statisztikai módszereket szűrésre alkalmazzuk: a *unithood* érték mérésére a C és NC értékeket, a *termhood* mérése a *weirdness* arány segítségével valósul meg.

1 Bevezetés

A terminológiakivonatolás (a továbbiakban TE) során a TE-alkalmazás egy adott, írott nyers szövegből annak terminusjelöltjeivel tér vissza. A terminusjelöltek kinyerése és szűrése történhet szabályalapú és statisztikai módszerekkel. A leggyakrabban használt módszer ezek kombinációja, a hibrid módszer, ami [2] és [6] szerint először a terminusjelöltek kinyerésére a statisztikát alkalmazzák, majd azok szűrésére nyelvi filtereket.

2 Korpusz

Korpusznak négy, francia nyelvű informatikai témájú szabadalom leírását választottuk, amelyekben 854 különböző terminus található; a szövegek átlagosan 3500 tokennel rendelkeznek. A négy szabadalmat kézileg is annotáltuk: bejelöltük bennük az összes terminust.

3 Módszer

A vizsgálatunk során a terminusok kinyerésekor a megszokottól eltérően fordított sorrendet alkalmazunk: a terminusjelölt-listát szabályalapú módszerekkel nyerjük ki, majd ezt különböző szűrőkkel szűrjük a pontosság növelése érdekében. A szabályalapú kinyeréshez és szűréshez szükség van a szöveg előfeldolgozására, amelyhez a szöveget mondatokra, majd tokenekre bontó, illetve azokat szófaji címkékkel ellátó *Machinese*-t [3] használtuk.

3.1 Terminusjelölt-lista létrehozása és első szűrése szabályalapú módszerekkel

A terminusjelöltek listájának kinyeréséhez a leggyakoribb mintákból (pl. főnév+főnév, főnév+prepozíció+főnév) véges állapotú automatát hozunk létre. Ezt az automatát illesztjük a már szófaji címkékkel ellátott szövegre. Az *és/vagy* típusú koordinációkat visszaállítjuk az eredeti alakjukra. Az így kapott mintákat szűrjük egy *stopword*-listával, ami a leggyakoribb (főnevet is tartalmazó) kifejezéseket szűri ki a szövegből, hogy azok ne kerülhessenek be a terminusjelölt-listába. Ilyen típusú szerkezetek a *par exemple* 'például', *en effet* 'ugyanis' stb. A tulajdonneveket pedig az OpenCalais projekt keretében létrehozott *OpenCalais Web Service API* [8] nevű alkalmazással szűrjük.

3.2 Terminusjelölt-lista szűrése statisztikai módszerekkel

A terminusjelöltekre a C és NC [5], *weirdness* [1] értékek kiszámítására szolgáló algoritmust alkalmazzuk. Mindhárom értékre igaz az, hogy minél nagyobb egy adott terminusnál annak értéke, annál valószínűbb, hogy az adott jelölt ténylegesen terminus.

A C-érték egy *unithood* mérték, ami azt mutatja meg, hogy egy adott terminusjelölt gyakrabban fordul-e elő önmagában vagy egy nagyobb egység részeként. Így például kiszűrhetők azok a melléknévi utómódosítók, amelyek az adott terminusnak nem lehetnek részei, mert a terminus részét nem képező melléknévi utómódosító és a főnévi fej közötti kohéziós érték alacsony lesz, ha ritkán fordulnak elő együtt.

Az NC-érték azt vizsgálja meg, hogy az adott terminusjelölt környezetében lévő szavak milyen valószínűséggel jelzik azt, hogy előttük vagy mögöttük terminus áll.

A *weirdness* pedig egy olyan *termhood* mérték, amely azt mutatja meg, hogy az adott terminusjelölt az adott szakszövegben vagy egy általános nyelvű korpuszban fordul elő gyakrabban. Ehhez egy általános keresőmotort használunk, az Exalead vállalat online keresőjét [4]: a saját alkalmazásunk minden egyes terminusjelöltnél lekérdezi annak gyakoriságát egy köznyelvi újság, a Le Figaro weboldaláról [7]. A Le Figaro keresési feltételként történő megadásának célja, hogy a keresőmotor ne kereshessen bárhol, hiszen így szakmai szövegekben is keresne, amit el kell kerülni.

A fent említett mértékeket először külön-külön alkalmazzuk az adott korpuszra, és megnézzük, hogy milyen hatékonyság érhető el ezeknél. Megkeressük minden változónál azt a határértéket, amely felett a legjobb a pontosság, fedés, illetve F-érték. Ezt követően egy összevont értéket is alkalmazunk, ami minden érték együttes eredményét veszi alapul.

4 Eredmények

A terminológiakivonatolás esetén a fedés a helyesen kinyert terminusok számának és az adott szövegben lévő terminusok számának a hányadosa, a pontosság a helyesen kinyert terminusok és az összes kinyert terminusok számának hányadosa, az F-érték pedig a fedés és pontosság szorzatának duplája osztva a fedés és a pontosság összegével [2].

A tisztán szabályalapú algoritmussal, tehát a mintákkal, körülbelül 0,78-as fedést és 0,59 értékű pontosságot (F-érték: 0,67) érhetünk el. A fedés és a pontosság értékei már akkor jelentősen nőnek, ha a mintaillesztés után a terminusjelölteket szűrjük az előre megadott, főnevet is tartalmazó fordulatokkal, valamint a benne szereplő tulajdonnevekkel: ekkor a pontosság 0,66 a fedés 0,83 (F-érték: 0,74). A statisztikai módszerek a várt eredményeket hozták: a pontosságot tudták növelni, de ezáltal a fedés csökkent. A legjobb pontosságot az általunk létrehozott kombinált érték biztosította, mely által ez az érték 0,89 lett. Az 1. táblázat foglalja össze az eredményeket a különböző algoritmusok esetén, ahol a legjobb értéket vastaggal emeltünk ki.

1. táblázat: Fedés, pontosság és F-érték a különböző módszerek esetén.

alkalmazott módszer	határérték	fedés	pontosság	F-érték
kinyerés mintákkal	-	0,7834	0,5895	0,6728
kinyerés mintákkal + szabályalapú szűrés	-	0,8285	0,6609	0,7353
weirdness				
	-	**0,8285**	0,6609	0,7353
	> 0,2595	0,7109	**0,6901**	0,7003
	> 0,0011	0,8285	0,6626	**0,7363**
C-érték:				
	-	**0,8285**	0,6609	0,7353
	> 2,8074	0,4574	**0,6917**	0,5506
	> -6,3399	0,8274	0,6618	**0,7354**
NC-érték:				
	-	**0,8285**	0,6609	**0,7353**
	> 1,5388	0,5620	**0,7098**	0,6273
C-NC érték				
	-	**0,8285**	0,6609	0,7353
	> 2,3807	0,4682	**0,6922**	0,5586
	> -4,8251	0,8274	0,6618	**0,7354**
kombinált érték				
	-	**0,8285**	0,6609	0,7353
	> 0,8468	0,0701	**0,8904**	0,13
	> 0,0867	0,8123	0,6759	**0,7379**

Bibliográfia

1. Ahmad, K., Gillam, L., Tostevin, L. Weirdness indexing for logical document extrapolation and retrieval (wilder). In: The Eighth Text REtrieval Conference (TREC-8). (1999) 717–724
2. Cabré, M. T., Bagot, R.E., Vivaldi Palatresi, J.: Automatic term detection. A review of current systems. In: Bourrigault, D., Jacquemin, Ch., L'Homme, M-C. (szerk.): Recent advantages in Computational Terminology. John Benjamins Publishing Co., Amsterdam/Philadelphia (2001) 53–87
3. Connexor – Technology – Machinese – Demo, http://www.connexor.eu/technology/machinese/demo/
4. Exalead search, http://www.exalead.com/search
5. Frantzi, K. T., Ananiadou, S.: The c/nc value domain independent method for multi-word term extraction. Journal of Natural Language Processing Vol. 6, No. 3 (1999) 145–179
6. Ha, L.A., Fernandez, G., Mitkov, R., Corpas, G.: Mutual bilingual terminology extraction. In: Calzolari, N. et al. (szerk.): Proceedings of LREC 2008 (CD-ROM). ELRA, Marrakech (2008) 1818–1824
7. Le Figaro online, http://www.lefigaro.fr/
8. OpenCalais Web Service API, http://www.opencalais.com/documentation/calais-web-service-api