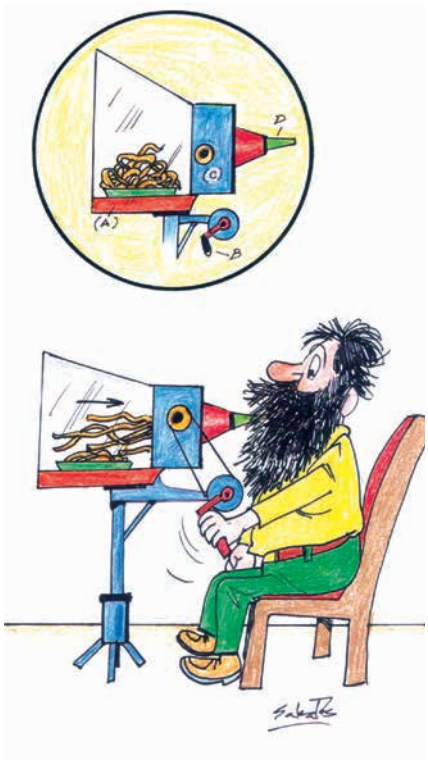
Terminologia per als invents i les patents

dossier

Processament automàtic de patents: un exercici de terminologia computacional

Leo Wanner
ICREA i Grup de Recerca en Tractament Automàtic del Llenguatge Natural (TALN)
Universitat Pompeu Fabra



Font: Col·lecció ENGINYERS BCN

Amb la importància creixent de la protecció de la propietat intel·lectual, les patents, l'únic mitjà legal disponible per documentar en detall una invenció i certificar-ne l'autor i la propietat intel·lectuals, han esdevingut un dels gèneres més destacats del discurs especialitzat.[1] L'aspecte negatiu d'aquest fet és que la revisió acurada i àgil de les sol·licituds i el seguiment de les infraccions de patents han esdevingut un repte extraordinari. Per afrontar aquest repte, tradicionalment s'han utilitzat els cercadors de patents. Tanmateix, a fi de no perdre cap material rellevant, aquests cercadors acostumen a basar-se en l'exhaustivitat (*recall-oriented*), fet que implica que els examinadors de patents i els especialistes en seguiment de patents han de revisar grans quantitats de material diàriament. L'únic instrument possible per alleugerir la seva càrrega de treball són les aplicacions de processament del llenguatge natural adaptades al gènere de les patents. Les aplicacions principals inclouen, com a mínim: *a*) el resum automàtic de patents; *b*) l'anàlisi (semàntica) de patents; *c*) la identificació de cadenes lèxiques, i *d*) la resolució de la correferència lèxica. Els resums de patents i els resultats de l'anàlisi semàntica de patents, que es poden reflectir en gràfics composicionals i funcionals, faciliten una visió general ràpida del contingut d'una patent sense necessitat de llegir-ne tot el document,

Terminàlia 16 (2017): 54-56 · DOI: 10.2436/20.2503.01.113
ISSN: 2013-6692 (impresa); 2013-6706 (electrònica) · *http://terminalia.iec.cat*

i la identificació de cadenes lèxiques i la resolució de la correferència lèxica destaquen els elements més importants d'una patent, cosa que, novament, fa que el procés de revisió sigui més ràpid. En el nostre treball abordem aquestes quatre tasques.

Fins ara, bona part de la recerca sobre el resum automàtic de patents s'ha centrat en el resum de les reivindicacions de patents (cf., p. ex., Shinmori *et al.*, 2003; Bouayad-Agha *et al.*, 2009; Trappey *et al.*, 2009), que presumiblement constitueixen la part central d'una sol·licitud de patent. Tanmateix, les altres parts (com ara la descripció) aprofundeixen, per exemple, en la realització preferida de la invenció i les seves possibles aplicacions, de manera que també és important resumir-les juntament amb les reivindicacions a fi d'obtenir un resum coherent de tot el document. A Codina-Filbà *et al.* (2017) presentem una proposta en aquest sentit.

La majoria de treballs que aborden el problema de l'anàlisi de patents fan referència a l'àmbit lèxic (Cascini *et al.*, 2009; Choi *et al.*, 2012; Xu *et al.*, 2015). Tanmateix, és important assenyalar que el grau d'abstracció del vocabulari varia significativament entre les reivindicacions i la descripció. Així doncs, una reivindicació pot recórrer als termes *aparell* i *dispositiu d'enregistrament*, mentre que en la descripció es pot fer referència a la mateixa invenció amb el terme *magnetòfon*. A fi d'obtenir una anàlisi acurada sobre la composició i el funcionament de la invenció patentada, hem de ser capaços d'identificar els components (o conceptes) a què fa referència cada terme. És a dir, hem de ser capaços d'obtenir la representació conceptual de la invenció a partir de la patent. Per tal d'assegurar que aquesta representació és suficientment flexible i es pot verificar formalment, fem servir les representacions del llenguatge d'ontologies web (OWL) (vegeu Dasiopoulou *et al.*, 2015).

Per norma general, els components de la invenció s'esmenten diverses vegades al llarg de les patents. La freqüència i la posició de les mencions indiquen la rellevància del component en qüestió. Per tant, és important recollir les «cadenes» dels termes que es refereixen al mateix component. Tal com ja s'ha apuntat abans, es fan servir diferents termes per referir-se al mateix component (cf., *aparell* i *magnetòfon*, *dispositiu* i *projector*, *font de llum* i *díode*, etc.), cosa que fa aquesta tasca encara més complexa. Els resultats del nostre treball en aquest àmbit estan documentats (p. ex., a Bouayad-Agha *et al.*, 2014; Brügmann *et al.*, 2015).

La resolució de la correferència lèxica a les patents està directament relacionada amb els problemes de l'anàlisi de patents i la identificació de cadenes lèxiques. En el discurs general preval la correferència amb un únic antecedent, com en l'exemple següent:

> 1) [*Una excavadora*]$_i$ *en* [*què*]$_i$ [*una estructura inferior mòbil*]$_j$ *està dotada d'un element superior giratori* [*en aquesta*]$_j$, *i*...

A les patents, però, també és molt comuna la correferència amb diversos antecedents:

> 2) *El circuit elèctric en què cadascun dels convertidors de corrent continu a corrent altern consta d'*[*un primer interruptor*]$_i$ [...] *i d'*[*un segon interruptor*]$_j$ [...]. *El circuit elèctric en què* [*el primer i el segon interruptor*]$_{i+j}$...

Cal resoldre els dos tipus de correferència per obtenir una visió objectiva de la importància d'un component. Per abordar aquest problema, adaptem l'estratègia del *multi-pass sieve* proposada per Raghunathan *et al.* (2010) (vegeu Bouayad-Agha *et al.*, 2014; Burga *et al.*, 2016).

És evident que en les quatre tasques esmentades anteriorment la terminologia és cabdal: per al resum de patents, la mètrica de rellevància que identifica els segments rellevants del document de patent que s'han d'incloure en el resum es basa en la correcta identificació i vinculació dels termes, mentre que l'anàlisi semàntica de patents, la identificació de cadenes lèxiques i la resolució de la correferència lèxica es basen en els termes per la seva mateixa naturalesa. Novament, cal tenir en compte les particularitats que caracteritzen la terminologia de les patents: *a*) els termes abstractes (com ara *aparell*, *dispositiu* o *mitjà*) actuen com a referents de significants més específics del mateix concepte; *b*) els termes que es fan servir a les patents consten molt sovint de diverses paraules (cf. *energia renovable*, *molí de vent*, *dispositiu electrònic recarregable*, etc.); *c*) en general, necessitem identificar no només termes concrets (és a dir, termes que denoten objectes concrets), sinó també termes predicatius (San Martin i L'Homme, 2014) que denoten accions i procediments.

Bibliografia

Bouayad-Agha, Nadjet; Burga, Alicia; Casamayor, Gerard; Codina, Joan; Nazar, Rogelio; Wanner, Leo (2014). «An exercise in reuse of resources: adapting general discourse coreference resolution for detecting lexical chains in patent documentation». A: *Proceedings of the Language Resources and Evaluation Conference (LREC)*, p. 3214-3221.

Bouayad-Agha, Nadjet; Casamayor, Gerard; Ferraro, Gabriela; Mille, Simon; Vidal, Vanesa; Wanner, Leo (2009). «Improving the comprehension of legal documentation: the case of patent claims». A: *Proceedings of the International Conference on Artificial Intelligence in Law*, p. 78-87.

Brügmann, Sören; Bouayad-Agha, Nadjet; Burga, Alicia; Carrascosa, Serguei; Ciaramella, Alberto; Ciaramella, Marco; Codina-Filbà, Joan; Escorsa, Enric; Judea, Alex; Mille, Simon; Müller, Andreas; Saggion, Horacio; Ziering, Patrick; Schütze, Hinrich; Wanner, Leo (2015). «Towards content-oriented patent document processing: intelligent patent analysis and summarization». *World Patent Information Journal*, vol. 40, p. 30–42.

Burga, Alicia; Cajal, Sergio; Codina-Filbà, Joan; Wanner, Leo (2016). «Towards multiple antecedent coreference resolution in specialized discourse». A: *Proceedings of the Language Resources and Evaluation Conference (LREC)*, p.2052-2057.

Cascini, Gaetano; Cugini, Umberto; Frillici, Francesco Saverio; Rotini, Federico (2009). «Computer-aided conceptual design through TRIZ-based manipulation of topological optimizations». A: *Proceedings of the 19th CIRP Design Conference–Competitive Design*.p. 263-271

Choi, Sungchol; Park, Hyunseok; Kang, Dongwoo; Lee, Jae Yeol L.; Kim, Kwangsoo (2012). «An SAO-based text mining approach to building a technology tree for technology planning». *Expert Systems with Applications*, vol. 39, núm. 13, p. 11443–11455.

Codina-Filbà, Joan; Bouayad-Agha, Nadjet; Burga, Alicia; Casamayor, Gerard; Mille, Simon; Müller, Andreas; Saggion, Horacio; Wanner, Leo (2017). «Using genre-specific features for patent summaries». *Information Processing and Management*, vol. 53, núm. 1, p. 151–174.

Dasiopoulou, Stamatia; Lohmann, Steffen; Codina, Joan; Wanner, Leo (2015). «Representing and visualizing text as ontologies: a case from the patent domain». A: *Proceedings of the Visualizations and User Interfaces for Ontologies and Linked Data (VOILA!) Workshop*, en el marc de la ISWC 2015 (Bethlehem, Pennsilvània), p. 83-90.

Raghunathan, Karthik; Lee, Heeyoung.; Rangarajan, Sudarshan; Chambers, Nathanael; Surdeanu, Mihai; Jurafsky, Daniel; Manning, Christopher. D. (2010). «A multi-pass sieve for coreference resolution». A: *Proceedings of the 2010 EMNLP*, p. 492–501.

San Martin, Antonio; L'Homme, Marie-Claude (2014). «Definition patterns for predicative terms in specialized dictionaries». A: *Proceedings of the Language Resources and Evaluation Conference (LREC)*, p 3748-3755.

Shinmori, Akihiro; Okumura, Manabu; Marukawa, Yuzo; Iwayama, Makoto (2003). «Patent claim processing for readability: structure analysis and term explanation». A: *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing*, p. 56–65.

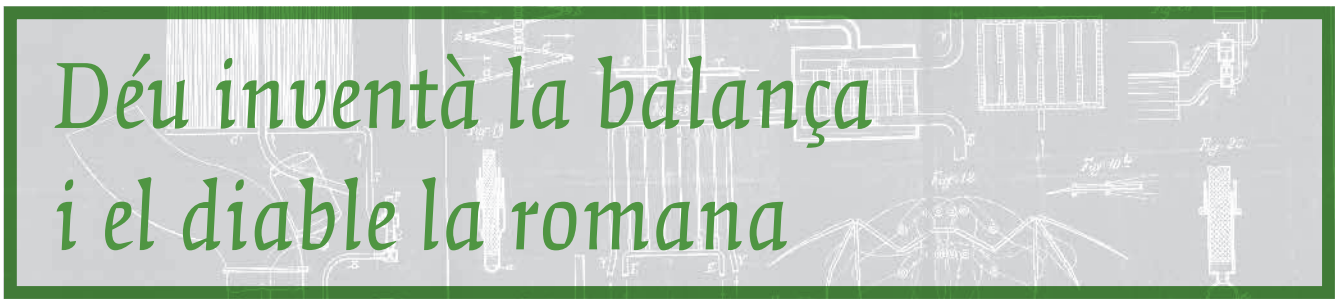
Trappey, Amy J. C.; Trappey, Charles V.; Wu, Chun-Yi (2009). «Automatic patent document summarization for collaborative knowledge systems and services». *Journal of Systems Science and Systems Engineering*, vol. 18, núm. 1, p. 71-94.

Xu, Hongjiao; Gui, Jie; Qu, Peng; Zhu, Xiaohua; Wang, Lijun (2015). «Exploring similarity between academic paper and patent based on Latent Semantic Analysis and Vector Space Model». A: *Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, p. 801-805.

Nota

1. Només el 2015, l'Oficina Europea de Patents va rebre unes 280.000 sol·licituds de patents; cf. *https://www.epo.org/about-us/annual-reports-statistics/statistics.html*.

El refrany



Font: *Paremiologia catalana. www.dites.cat*