Terminology extraction

Terminology extraction (also known as **term** extraction, **glossary** extraction, term **recognition**, or terminology **mining**) is a subtask of information extraction. The goal of terminology extraction is to automatically extract relevant terms from a given corpus.[1]

In the semantic web era, a growing number of communities and networked enterprises started to access and interoperate through the internet. Modeling these communities and their information needs is important for several web applications, like topic-driven web crawlers,[2] web services,[3] recommender systems,[4] etc. The development of terminology extraction is also essential to the language industry.

One of the first steps to model the knowledge domain of a virtual community is to collect a vocabulary of domain-relevant terms, constituting the linguistic surface manifestation of domain concepts. Several methods to automatically extract technical terms from domain-specific document warehouses have been described in the literature.[5][6][7][8][9][10][11][12][13][14][15][16][17]

Typically, approaches to automatic term extraction make use of linguistic processors (part of speech tagging, phrase chunking) to extract terminological candidates, i.e. syntactically plausible terminological noun phrases, NPs (e.g. compounds "credit card", adjective-NPs "local tourist information office", and prepositional-NPs "board of directors" - in English, the first two constructs are the most frequent[18]). Terminological entries are then filtered from the candidate list using statistical and machine learning methods. Once filtered, because of their low ambiguity and high specificity, these terms are particularly useful for conceptualizing a knowledge domain or for supporting the creation of a domain ontology or a terminology base. Furthermore, terminology extraction is a very useful starting point for semantic similarity, knowledge management, human translation and machine translation, etc.

Bilingual terminology extraction

The methods for terminology extraction can be applied to parallel corpora. Combined with e.g. co-occurrence statistics, candidates for term translations can be obtained.[19] Bilingual terminology can be extracted also from comparable corpora[20] (corpora containing texts within the same text type, domain but not translations of documents between each other).

See also

- Computational linguistics
- Glossary
- Natural language processing
- Domain ontology
- Subject indexing
- Taxonomy (general)
- Terminology
- Text mining
- Text simplification

References

1. Alrehamy, Hassan H; Walker, Coral (2018). "SemCluster: Unsupervised Automatic Keyphrase Extraction Using Affinity Propagation". *Advances in Computational Intelligence Systems*. Advances in Intelligent Systems and Computing. **650**. pp. 222–235. doi:10.1007/978-3-319-66939-7_19 (https://doi.org/10.1007/978-3-319-66939-7_19). ISBN 978-3-319-66938-0.
2. Menczer F., Pant G. and Srinivasan P. Topic-Driven Crawlers: machine learning issues (http://citeseer.ist.psu.edu/menczer02topicdriven.html)

3. Fan J. and Kambhampati S. A Snapshot of Public Web Services (http://portal.acm.org/citation.cfm?id=1058150.1058156), in ACM SIGMOD Recordarchive Volume 34 , Issue 1 (March 2005).
4. Yan Zheng Wei, Luc Moreau, Nicholas R. Jennings. A market-based approach to recommender systems (http://portal.acm.org/citation.cfm?id=1080344&dl=ACM&coll=&CFID=15151515&CFTOKEN=6184618), in ACM Transactions on Information Systems (TOIS), 23(3), 2005.
5. Bourigault D. and Jacquemin C. Term Extraction+Term Clustering: an integrated platform for computer-aided terminology (http://acl.ldc.upenn.edu/E/E99/E99-1003.pdf) Archived (https://web.archive.org/web/20060619123604/http://acl.ldc.upenn.edu/E/E99/E99-1003.pdf) 2006-06-19 at the Wayback Machine., in Proc. of EACL, 1999.
6. Collier, N.; Nobata, C.; Tsujii, J. (2002). "Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain". *Terminology*. **7** (2): 239–257. doi:10.1075/term.7.2.07col (https://doi.org/10.1075/term.7.2.07col)
7. K. Frantzi, S. Ananiadou and H. Mima. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. (http://arnetminer.org/viewpub.do?pid=986374) In: C. Nikolau and C. Stephanidis (Eds.) International Journal on Digital Libraries, Vol. 3, No. 2., pp. 115-130.
8. K. Frantzi, S. Ananiadou and J. Tsujii. (1998) The C-value/NC-value Method of Automatic Recognition of Multi-word Terms (http://dl.acm.org/citation.cfm?id=696825) In: ECDL '98 Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries, pp. 585-604. ISBN 3-540-65101-2
9. L. Kozakov; Y. Park; T. Fin; Y. Drissi; Y. Doganata & T. Cofino. (2004). "Glossary extraction and utilization in the information search and delivery system for IBM Technical Support" (http://www.research.ibm.com/people/y/yurdaer/papers/ibmsysjournal2004a.pdf) (PDF). *IBM Systems Journal* **43** (3).
10. Navigli R. and Velardi, P. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites (http://portal.acm.org/citation.cfm?id=1105712) Computational Linguistics. 30 (2), MIT Press, 2004, pp. 151-179
11. Oliver, A. and Vàzquez, M. TBXTools: A Free, Fast and Flexible Tool for Automatic Terminology Extraction (http://aclweb.org/anthology/R15-1062) Proceedings of Recent Advances in Natural Language Processing (RANLP 2015), 2015, pp. 473–479
12. Y. Park, R. J. Byrd, B. Boguraev "Automatic glossary extraction: beyond terminology identification" (http://portal.acm.org/citation.cfm?id=1072370&dl=ACM&coll=&CFID=15151515&CFTOKEN=6184618), International Conference On Computational Linguistics, Proceedings of the 19th international conference on Computational linguistics aipei, Taiwan, 2002.
13. Sclano, F. (https://web.archive.org/web/20070106074102/http://lcl2.di.uniroma1.it/~sclano/) and Velardi, P. (http://www.dsi.uniroma1.it/~velardi/welcome.htm) TermExtractor (https://web.archive.org/web/20070214092612/http://lcl2.di.uniroma1.it/termextractor/) a Web Application to Learn the Shared Terminology of Emergent Web Communities. To appear in Proc. of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007). Funchal (Madeira Island), Portugal, March 28–30th, 2007.
14. P. Velardi, R. Navigli, P. D'Amadio. Mining the Web to Create Specialized Glossaries (http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4629722) IEEE Intelligent Systems, 23(5), IEEE Press, 2008, pp. 18-25.
15. Wermter J. and Hahn U. Finding New terminology in Very large Corpora (http://portal.acm.org/citation.cfm?id=1088648), in Proc. of K-CAP'05, October 2–5, 2005, Banff, Alberta, Canada
16. Wong, W., Liu, W. & Bennamoun, M. (2007) Determining Termhood for Learning Domain Ontologies using Domain Prevalence and Tendency (http://portal.acm.org/citation.cfm?id=1378245.1378253) In: 6th Australasian Conference on Data Mining (AusDM); Gold Coast. ISBN 978-1-920682-51-4
17. Wong, W., Liu, W. & Bennamoun, M. (2007) Determining Termhood for Learning Domain Ontologies in a Probabilistic Framework (http://portal.acm.org/citation.cfm?id=1378254&jmp=cit&coll=GUIDE&dl=GUIDE) In: 6th Australasian Conference on Data Mining (AusDM); Gold Coast. ISBN 978-1-920682-51-4
18. Alrehamy, Hassan H; Walker, Coral (2018). "SemCluster: Unsupervised Automatic Keyphrase Extraction Using Affinity Propagation". *Advances in Computational Intelligence Systems* Advances in Intelligent Systems and Computing. **650**. pp. 222–235. doi:10.1007/978-3-319-66939-7_19 (https://doi.org/10.1007/978-3-319-66939-7_19) ISBN 978-3-319-66938-0
19. Macken, Lieve, Els Lefever, and Veronique Hoste. "TExSIS: Bilingual terminology extraction from parallel corpora using chunk-based alignment." Terminology 19.1 (2013): 1-30.
20. Sharoff, Serge; Rapp, Reinhard; Zweigenbaum, Pierre; Fung, Pascale (2013), *Building and Using Comparable Corpora* (https://www.springer.com/cda/content/document/cda_downloaddocument/9783642201277-c1.pdf?SGWID=0-0-45-1442068-p174109864) (PDF), Berlin: Springer-Verlag