KD Strikes Back: from Keyphrases to Labelled Domains Using External Knowledge Sources

Giovanni Moretti[1], Rachele Sprugnoli[1-2], Sara Tonelli[1]
[1]Fondazione Bruno Kessler, Trento
[2] Università di Trento
{moretti,sprugnoli,satonelli}@fbk.eu

Abstract

English. This paper presents L-KD, a tool that relies on available linguistic and knowledge resources to perform keyphrase clustering and labelling. The aim of L-KD is to help finding and tracing themes in English and Italian text data, represented by groups of keyphrases and associated domains. We perform an evaluation of the top-ranked domains using the 20 Newsgroup dataset, and we show that 8 domains out of 10 match with manually assigned labels. This confirms the good accuracy of this approach, which does not require supervision.

Italiano. *In questo lavoro descriviamo L-KD, un sistema che utilizza risorse linguistiche e basate su conoscenza per raggruppare concetti-chiave e categorizzarli. L'obiettivo di L-KD è quello di supportare gli utenti nel rilevare la presenza di specifici temi in documenti italiani e inglesi, rappresentandoli attraverso gruppi di concetti-chiave e relativi domini. Abbiamo valutato l'affidabilità del sistema analizzando i domini più rilevanti nel 20 Newsgroup dataset, e dimostrando che 8 su 10 domini nel gold standard sono assegnati correttamente anche dal sistema. Questa valutazione conferma le buone performance di L-KD, senza il bisogno di supervisione.*

1 Introduction

With the increasing availability of large document collections in digital format, companies, organizations but also non-expert users face everyday the need to efficiently extract and categorize relevant information from large corpora. The possibility to extract key-concepts and assign them to a domain without the need of supervision would allow them to systematically track the flow of information and retain only relevant content at two granularity levels: key-concepts, and domains to which these key-concepts can be ascribed. Although topic models (Blei et al., 2003) can be used to this purpose, they have two main drawbacks: the number of topics for a corpus is arbitrary and topics are often not labelled.

In this work, we present a solution to the aforementioned research problem by presenting L-KD (Labelled-KD), a tool to perform keyphrase clustering and labelling through the exploitation of external linguistic and knowledge resources. The tool takes advantage of the availability of Keyphrase Digger[1] (KD), a multilingual rule-based system that detects a weighted list of n-grams representing the most important concepts in a text (Moretti et al., 2015). These key-concepts are then linked to WordNet Domains (Magnini and Cavaglia, 2000) in order to create clusters of key-concepts labelled by domain. The problem of ambiguous concepts, i.e. possibly belonging to more than one WordNet domain, is tackled by using ConceptNet 5 (Speer and Havasi, 2013), a multilingual knowledge source containing single and multi-word concepts linked to each other by a broad set of relations covering different types of associations. The outcome of this study is the L-KD tool, supporting both English and Italian, which we make available to the research community[2]. L-KD takes in input a document in plain text format, and outputs the ranked list of semantic domains discussed in the documents, each associated with a set of keyphrases.

[1]http://dh.fbk.eu/technologies/kd
[2]https://dh.fbk.eu/technologies/l-kd

2 Related Works

In the last years, a number of works dealing with the unsupervised clustering of keyphrases has been presented (Hasan and Ng, 2014). Liu et al. (2009) use Wikipedia and co-occurrence-based statistics to semantically cluster similar keyphrases in a set of unweighted topics. In order to improve this approach by weighting topics, Liu et al. (2010) and Grineva et al. (2009) propose a topic-decomposed PageRank and a network analysis algorithm respectively to perform hierarchical clustering. Our method is simpler than the previously mentioned studies, and relies on available resources to label the clusters. Indeed, the lists of terms listed in the topics are not always easy to interpret (Aletras et al., 2015), and adding a label that captures the meaning of each cluster is a way to enhance its understanding. The problem of interpretation affects also the output of topic modelling algorithms, i.e. unsupervised statistical methods such as Latent Dirichlet Allocation (Blei et al., 2003). Many techniques have been developed to automatically label topics for example by using probabilistic approaches (Mei et al., 2007), Wikipedia links (Xu and Oard, 2011) and DBpedia structured data (Hulpus et al., 2013). As for the automatic labelling of keyphrase clusters, Carmel et al. (2009) adopt Wikipedia as an external resource to extract candidate labels. To the best of our knowledge, no available system performs this task by combining WordNet Domains and ConceptNet 5.

3 System Overview

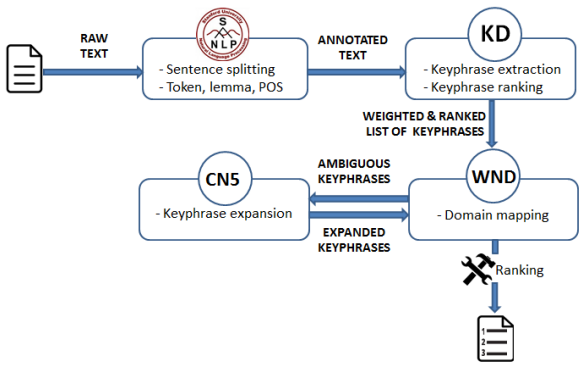

Figure 1: General workflow underlying L-KD for English documents with the steps involving the use of Stanford CoreNLP, KD, WordNet Domains (WND) and ConceptNet 5 (CN5).

```
nature -> [country_side, environment,
flower_lake, mother_goddess, science,
outdoor_world, tree_animal, many_wonder,
great_place, fauna, flora, ecology,
conservation, country, ecosystem,
everything_live, flower, plant …….]
```

```
{Biology=19, Geography=13, Plants=8,
Animals=5 …}
```

Figure 2: Excerpt of the expansion of an ambiguous keyphrase using ConceptNet 5 (top) and top domains assigned to this expansion (bottom).

L-KD performs several steps (see Fig. 1) to semantically cluster keyphrases and label each cluster:

1) Text preprocessing: Stanford CoreNLP (Manning et al., 2014) is used to split sentences, tokenize, lemmatize and tag the part-of-speech of the input English text. For Italian texts, we rely on Tint[3], a suite of NLP tools (Aprosio and Moretti, 2016) based on the Stanford CoreNLP pipeline.

2) Keyphrase extraction and ranking: L-KD integrates KD, a keyphrase extraction tool that combines statistical and linguistic knowledge, given by recurrent relevant PoS patterns, to extract single words and multi-token expressions encoding the main concepts of a document. A detailed description of KD functionalities is given in Moretti et al. (2015). The output of this step is a weighted and ranked list of keyphrases.

3) Domain mapping: L-KD maps the lemma forms of keyphrases with the lemmas in WND aligned to WordNet 3.0[4]. For Italian we rely on the data available through the Open Multilingual WordNet project (Bond and Paik, 2012) as a bridge between lemmas and WND. In case of multi-token expressions (e.g. "federal government"), the system looks for a perfect match. If no match is found, the tokens are splitted and only the nouns are searched in WND (e.g. "government"). A list of domain-keyphrases associations is created, as well as a list of ambiguous keyphrases. The latter comprises those that are assigned to the `Factotum` domain and those that could belong to several domains, if none of them contains > 3 keyphrases. This threshold was manually set in order to identify domains that are likely to be little relevant.

4) Expansion of ambiguous keyphrases: The lemmas of ambiguous keyphrases are aligned with

[3] `http://tint.fbk.eu/`

[4] Courtesy of Carlo Strapparava.

the lemmas in ConceptNet 5 and are expanded by retrieving all the connected concepts following ConcepNet 5 relations. L-KD relies on a subset of relations including hierarchical (*HasA*, *PartOf*, *MadeOf*, *IsA*, *DerivedFrom*) and synonymous (*Synonym*, *RelatedTo*) ones (Mukherjee and Joshi, 2013). Functional relations such as *CapableOf* and *UsedFor* are not taken into consideration because the concepts evoked by these relations may be too far from the original meaning of the key-concept. The upper part of Fig. 2 shows how "nature", an ambiguous keyphrase, is expanded following this procedure. Examples of the relations that lead to this expansion are the following:
- nature ⇒ RelatedTo ⇒ flora
- nature ⇒ IsA ⇒ great_place
- nature ⇒ HasA ⇒ many_wonder

5) Domain mapping of expanded keyphrases: All the lemmas included in the expansion created in the previous step are mapped to domains using WND. The lower part of Fig. 2 reports the top domains related to the expansion of "nature" together with the number of lemmas associated with them, e.g. 19 lemmas are mapped to the `Biology` domain. A relevance score (i.e. number of keyphrases associated with a domain) is computed for the domains retrieved for each expanded keyphrase. Domains are then compared with the ones found in Step (3) starting from the domain with the highest score. If it is already present in the domain-keyphrases list compiled in Step (3), then the keyphrase is associated with this domain, otherwise the other domains are checked. If the domain is not present in the list, it is added to the list with its associated keyphrase. The final relevance score of the domains is recalculated at the end of this step. Four sub-domains of `Factotum`, i.e. `Time_Period`, `Person`, `Metrology` and `Numbers`, which are very generic, usually have a high relevance because they tend to include many keywords. Therefore, we introduce a final reweighting step to deboost them.

6) Final ranking. L-KD creates a final ranked list of domains associated with clusters of keyphrases. The ranking is based on the relevance score of the domains as described in the previous step and on the rank of keyphrases as given by KD in step (2).

4 Evaluation

We evaluated L-KD using the 20 Newsgroup dataset (Joachims, 1996), a corpus of 20,000 documents extracted from UseNet discussion groups. This dataset is freely available online[5] and has been often employed to train and test text categorization algorithms (Moschitti and Basili, 2004). Specifically, each of its documents was manually assigned to one out of twenty different categories, which can be easily mapped to WND labels. Although L-KD can assign a ranked list of domains to one or more documents, thus providing a richer representation of the document(s) content, we did not find a suitable gold standard to evaluate the rank. Therefore, we limit our evaluation to the top-ranked domain extracted by the tool. We also decided to group Newsgroup categories that are strictly related to each other: e.g. documents in *talk.religion.misc*, *alt.atheism*, and *soc.religion.christian* all discuss religious issues and for this reason their texts are collapsed in a single category.
Table 1 reports the results of L-KD on the documents included in each category or group of categories. The second column shows the top two domains retrieved by the system and the third column presents some of the extracted keyphrases. Only in 2 cases out of 10, the first ranked domain does not perfectly match the original category: indeed `Law` is the top domain of *sci.eletronics* and of the documents related to political themes (*talk.politics.misc*, *talk.politics.guns*, *talk.politics.mideast*). We can notice that `Law` is a very frequent domain because it contains generic and recurring words such as "article", "opinion" and "information". In the rest of the cases (8 out of 10), the match between the first ranked domain and the original category is perfect: for example, the domain with the highest rank for documents discussing computer technologies is `Computer_Science`. In many cases also the second domain is extremely relevant. For instance, *misc.forsale* contains messages of people searching or selling goods with a focus on computer devices and components: the first retrieved domain is `Commerce` and the second one is `Computer_Science`. Each domain is associated with pertinent keyphrases such as "best offer" for the first domain and "floppy drive" for the second.

[5] `http://qwone.com/~jason/20Newsgroups/`

ORIGINAL CATEGORIES	TOP DOMAINS	KEYPHRASES
sci.med	**Medicine**	doctor, infectious disease, side effect
	School	course, science, study
sci.space	**Astronomy**	solar system, physical universe, satellite
	Transport	spacecraft, shuttle, high-speed collision
sci.crypt	**Computer_Science**	internet, e-mail, bit
	Law	security, second amendment, criminal
sci.electronics	Law	article, opinion, information
	Electricity	amateur radio, voltage, wire
talk.religion.misc - alt.atheism - soc.religion.christian	**Religion**	christian, atheist, objective morality
	Law	law, evidence, private activities
rec.sport.baseball - rec.sport.hockey	**Sport**	game, playoff, second period
	Play	player, baseball
rec.autos - rec.motorcycles	**Transport**	car, mph, front wheel
	Law	article, opinion
comp.graphics - comp.os.mswindows.misc - comp.sys.ibm.pc.hardware - comp.windows.x - comp.sys.mac.hardware	**Computer_Science**	software, hard drive, anonymous ftp
	Publishing	article, opinion
talk.politics.misc - talk.politics.guns - talk.politics.mideast	Law	opinion, second amendment
	Transport	road, ways of escape
misc.forsale	**Commerce**	best offer, price, excellent condition
	Computer_Science	hard drive, floppy drive, email

Table 1: Results of L-KD on the 20 Newsgroup dataset. The original categories are compared with the top domains extracted by the systems. Examples of keyphrases are provided for each domain. Perfect matches between the main theme of the original classification and L-KD top domains are in bold.

5 Use Case: the De Gasperi Project

L-KD has been recently applied to the analysis of the complete corpus of public writings by Alcide De Gasperi (De Gasperi, 2006) in the context of a research project, whose goal is to give insight into De Gasperi's communication strategy with the help of innovative tools for text analysis. We processed the 2,762 documents (around 3,000,000 tokens) in the corpus, published between 1901 and 1954, to analyse which domains appeared in the collection and how they changed over time. The advantage of L-KD is that it can provide both a distant view, by computing aggregated information on the domains, and a close reading of the documents, showing which key-concepts are mapped to which domain. As an example, we report in Fig. 3 the analysis related to two documents, entitled "Rene de la Tour du Pin" and "I cattolici nell'evoluzione sociale'. For each of them, the dendogram shows the three top domains and the associated key-concepts. The proposed analysis was validated at different granularities by two history scholars, who confirmed the consistency of L-KD analysis and found correspondences between the top domains and relevant events in De Gasperi's life.

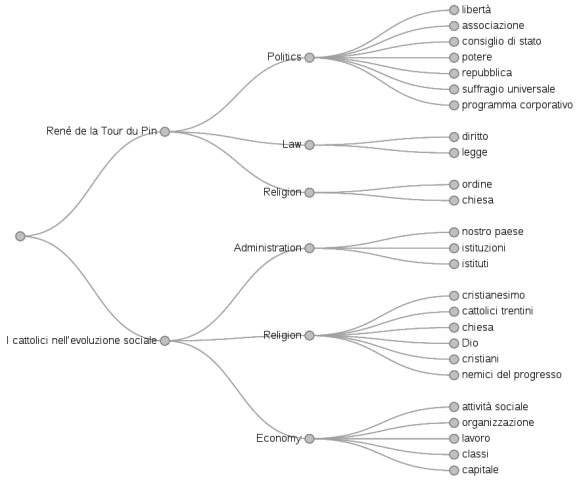

Figure 3: Dendogram related to two documents from De Gasperi's corpus

6 Tool Availability

L-KD is available as a web application[6] through which users can copy&paste a document and run the tool processing it on the fly. This application makes L-KD easily accessible also by users without a technical background.

[6] http://dhlab.fbk.eu:8080/L_KD/

In the application some parameters are given, while others can be changed by the user according to his/her needs. As for the fixed parameters, proper names are always discarded so to exclude them from the list of keyphrases: this setting is justified by the fact that WordNet, and consequently WND, contains few proper nouns[7] while we want to maximize the mapping. For the same reason, short keyphrases, i.e. single words and multi-token expressions with a maximum length of 4 words[8], are preferred. On the contrary, the minimum number of occurrences for a word or expression to be considered as a candidate keyphrase and the number of keyphrases to be extracted can be customized by the user. For example, in case of short documents, a low number of keyphrases (e.g. up to 20) can be set together with a minimum frequency of 1 or 2 (in a short text repetitions are less likely to occur). For long documents more keyphrases can be extracted: in this way it would be easy to find clusters covering multiple themes.

7 Conclusions and Future Works

This paper presents L-KD, a tool that extracts keyphrases from text data, clusters them according to the domain and assigns a label to each cluster. The process underlying L-KD is based on the exploitation of external linguistic and knowledge resources, i.e. WordNet Domains and ConceptNet 5. Our tool can process both English and Italian texts of different length and content, from a single news article to an entire book, from single-theme to multi-theme documents.

In the future we will explore different research directions. First of all we want to evaluate the tool on Italian data, even if we have not found a suitable gold standard so far. Resorting to crowd-sourcing may be a viable solution. We expect lower performances than the ones obtained for English, given that the current mapping between Open Multilingual WordNet and WordNet 3.0 covers only the 32.5% of the English synsets: this consequently affects the mapping on the domains of WND. Moreover, the coverage of Italian in ConcepNet 5 is limited. As for the availability of L-KD, we plan to release the tool as a standalone module. It will also be integrated in the ALCIDE platform (Moretti et al., 2016) that supports the analysis of large document collections for humanities studies.

[7] Only the 9.4% of synsets are tagged as being instances, i.e. proper nouns, in WordNet 3.0 (Abrate et al., 2012).

[8] In WordNet 3.0 only the 0.2% of noun synsets have a length greater than 4 words.

Acknowledgments

The research leading to this paper was partially supported by the EU Horizon 2020 Programme via the SIMPATICO Project (H2020-EURO-6-2015, n. 692819). We thanks Alessio Palmero Aprosio for his help in the evaluation process.

References

Matteo Abrate, Clara Bacciu, Andrea Marchetti, and Maurizio Tesconi. 2012. WordNet atlas: a web application for visualizing WordNet as a zoomable map. In *GWC 2012 6th International Global Wordnet Conference*, page 23.

Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. 2015. Evaluating topic representations for exploring document collections. *Journal of the Association for Information Science and Technology*.

Alessio Palmero Aprosio and Giovanni Moretti. 2016. Italy goes to Stanford: a collection of CoreNLP modules for Italian. *arXiv preprint arXiv:1609.06204*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Francis Bond and Kyonghee Paik. 2012. A survey of WordNets and their licenses. *Small*, 8(4):5.

David Carmel, Haggai Roitman, and Naama Zwerdling. 2009. Enhancing cluster labeling using wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 139–146. ACM.

A. De Gasperi. 2006. Scritti e discorsi politici. In E. Tonezzer, M. Bigaran, and M. Guiotto, editors, *Scritti e discorsi politici*, volume 1. Il Mulino.

Maria Grineva, Maxim Grinev, and Dmitry Lizorkin. 2009. Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th international conference on World wide web*, pages 661–670. ACM.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273. Association for Computational Linguistics.

Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. 2013. Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the*

sixth ACM international conference on Web search and data mining, pages 465–474. ACM.

Thorsten Joachims. 1996. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. Technical report, DTIC Document.

Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. 2009. Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 257–266. Association for Computational Linguistics.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 366–376. Association for Computational Linguistics.

Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating Subject Field Codes into WordNet. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499. ACM.

Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. 2015. Digging in the dirt: Extracting keyphrases from texts with kd. In *Proceedings of CLiC-it 2016*, page 198.

Giovanni Moretti, Rachele Sprugnoli, Stefano Menini, and Sara Tonelli. 2016. ALCIDE: Extracting and visualising content from large document collections to support Humanities studies. *Knowledge-Based Systems*, 111:100–112.

Alessandro Moschitti and Roberto Basili. 2004. Complex linguistic features for text classification: A comprehensive study. In *European Conference on Information Retrieval*, pages 181–196. Springer.

Subhabrata Mukherjee and Sachindra Joshi. 2013. Sentiment Aggregation using ConceptNet Ontology. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP)*, pages 570–578.

Robert Speer and Catherine Havasi. 2013. ConceptNet 5: A large semantic network for relational knowledge. In *The Peoples Web Meets NLP*, pages 161–176. Springer.

Tan Xu and Douglas W Oard. 2011. Wikipedia-based topic clustering for microblogs. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10.