Forecasting Emerging Trends from Scientific Literature

Kartik Asooja, Georgeta Bordea, Gabriela Vulcu, Paul Buitelaar
Insight Centre for Data Analytics
NUI Galway, Ireland
firstname.lastname@insight-centre.org

Abstract

Text analysis methods for the automatic identification of emerging technologies by analyzing the scientific publications, are gaining attention because of their socio-economic impact. The approaches so far have been mainly focused on retrospective analysis by mapping scientific topic evolution over time. We propose regression based approaches to predict future keyword distribution. The prediction is based on historical data of the keywords, which in our case, are LREC conference proceedings. Considering the insufficient number of data points available from LREC proceedings, we do not employ standard time series forecasting methods. We form a dataset by extracting the keywords from previous year proceedings and quantify their yearly relevance using tf-idf scores. This dataset additionally contains ranked lists of related keywords and experts for each keyword.

Keywords: Trend analysis, Regression, Emerging trends, Keyword extraction

1. Introduction

The rapidly changing landscape of scientific research calls for automated text analysis methods that reliably monitor and identify emerging trends from scientific publications. In recent years, emerging technologies draw an increased interest from research and development departments, academic publishing companies, and policy makers, because of their high potential for socio-economic impact. Another area where early identification of growing trends plays an import role is for organizing scientific dissemination events. A broad overview of emergent trends is needed for conference organisers to be able to identify hot topics for call for papers and recruit reviewers that are experts in these areas. In social science, an emergent technology is defined in terms of novelty, fast growth, coherence within a community of practice, and prominent impact (Rotolo et al., 2015). But a certain level of uncertainty and ambiguity is associated with emerging technologies, as their future impact is unclear and there is little consensus on their terminologies and meaning. The method proposed in this work models the growth pattern of a scientific topic to predict its distribution in the future. Although approaches based on topic models have received a lot of attention, they still require a considerable amount of manual work and expert knowledge to identify a collective topic for human-readable labels. Therefore, in this work we rely on previous work on keyphrase extraction to identify scientific topics (Bordea et al., 2013). Recent analysis of keyphrase creation dynamics shows that they are frequently used in scientific discourse to signal novelty (Adar and Datta, 2015).

Solutions based on citation analysis are less appropriate for contemporary analysis of emergent trends, as this type of data is less robust for recent documents and can not be applied as soon as documents become available. Where the majority of previous approaches apply retrospective analysis, by mapping the evolution of scientific topics over time (Zhou et al., 2006; He et al., 2009; Bolelli et al., 2009), in this paper, we address the problem of actually predicting temporal distribution of the keywords at some point in the future.

We also generate a dataset based on all of the LREC[1] conference proceedings. We use Saffron[2] framework to extract the keywords (or topics) and topic experts (Bordea et al., 2013). This dataset consists of the tf-idf score based temporal evolution of the different keywords in LREC conference, and related keywords and topic experts for each topic.

This paper is organized as follows. First, we discuss the related work on trend evolution in Section 2. Then, in Section 3, we provide an overview of the Saffron system which we use for generating the dataset. In section 4 we discuss our approach for forecasting the emerging trends. Then, we present evaluation of the approach with a description of the dataset in Section 5. Finally, we conclude the paper in Section 6.

2. Related Work

Previous work on trend evolution is mainly concerned with a historical analysis of the change in scientific topics over time, assuming that a user can manually spot emerging trends by investigating past dynamics of scientific topics. But this often requires considerable experience and understanding of the analyzed domain.

The related work in this area can be roughly categorized in two main directions. On one hand there are methods that perform topic-level analysis (Mörchen et al., 2008; He et al., 2009; Wang et al., 2008) starting from classical topic models such as LDA (Blei, 2012), followed by an analysis of how the topics change from one time period to the other; and more sophisticated topic models such as time dynamic topic models (Wang et al., 2008) where time is part of the model. On the other hand there are methods that investigate emerging trends using word frequency analysis (Adar and Datta, 2015; Guo et al., 2011; Lent et al., 1997) by tracking frequencies of keywords over time and creating trend-like

[1] http://lrec-conf.org/

[2] http://saffron.insight-centre.org/

patterns that users can query.
Our work is set in the second category, representing scientific topics using keywords, but in addition we focus on automatically predicting keyword distribution, rather than simply mapping the evolution of scientific topics over years. Compared to our approach, methods based on topic modelling or clustering require expert knowledge to manually label topics. To address this problem, previous works relied on predefined vocabularies (e.g., MESH), but these are not available for many scientific fields and may not cover most recent research topics.

3. The Saffron system

Saffron is a research framework, which provides services like keyword extraction, entity linking, taxonomy extraction, expertise mining, and data visualization. Saffron already provides an explorer[3] over the different research topics and experts in the domain of Natural Language Processing, Information Retrieval, and Semantic Web by analyzing the proceedings of conferences such as LREC, ACL, CLEF and ISWC. For our task, we require term extraction and their temporal growth for the past conferences. Term extraction is a central component in Saffron, with a focus on extracting terms of an intermediate level of specificity, which are useful for summarization and classification (Bordea et al., 2013).
Saffron provides an interesting trend analysis on the developments of the different topics in the research community in the previous years. For instance, Figure 2 shows the growth of the topic "Statistical Machine Translation" in the NLP community using the ACL anthology corpus[4]. However, Saffron does not provide any topic trend forecast for the coming years. Thus, in this paper, we propose a future trend forecast extension to Saffron based on regression methods.

4. Approach

The problem of predicting emerging trends can be ideally modeled in terms of forecasting a time series, which is a set of observations x_t, each one being recorded at a specific time *t* (De Gooijer and Hyndman, 2006). However, considering the sparse and insufficient number of data points available through biennial LREC proceedings (2000-2014) to model the keyword growth, we approach this problem using regression methods. In our case, the observation can be in the form of any numerical statistic such as tf-idf, which is a measure of the importance of a keyword computed from all the published papers in a particular period. In this way, our approach predicts future keyword distributions based on previously observed values. We present two alternative types of models for forecasting keyword distribution: fine-grained models that make a separate prediction for each keyword, and a global model that predicts the overall distribution of keywords.

[3]LREC on Saffron: http://saffron.insight-centre.org/lrec
[4]ACL Anthology corpus: http://aclweb.org/ anthology/

Individual Models for Keywords:
In this approach, we individually model the time series data for every keyword. However, as we use regression, we assume that the previous year values are independent of each other, which is not true for time series data. Also, in order to use regression here, we need to fix the number of independent variables unlike usual approaches for time series forecasting. So, we consider the last n years of keyword tf-idf data as the independent variables, where n is 4 for our experiments. We consider linear and polynomial prediction models to fit the growth/decay curves of the phrases. We use a least-squares fit to predict the curve, which minimises the sum of the squared errors in the data series.The slopes of the least squares lines give the "trend" values for the different keywords. However, fitting the growth curves with a line or a polynomial of specific degree for all the keywords is not optimal, as different keywords have different growth patterns. Therefore, we experiment with learning the polynomial degree of the growth curve for each keyword, as a variety of functions can be chosen for fitting the time series data.

Global Model for Keywords:
In this approach, we try to learn a single regression model which can approximate the whole data, rather than learning separate curve equations for each keyword. We form a dataset consisting of multiple instances of n observation values of any keyword and its $n + 1^{th}$ observation as the target value. We take n=4 for our experiments. This model essentially learns an averaged out curve which tries to fit the whole data.

5. Evaluation

We compare the above two regression based approaches to model the temporal development of scientific topics, and identify emerging trends for a popular conference in the domain of computational linguistics, the International Conference on Language Resources and Evaluation (LREC). In the presented experiments, we use the LREC proceedings for evaluation purposes, although this is a domain independent solution and is applicable to a larger collection of scientific publications from any research area.

5.1. Data

To begin with, we extract roughly 55K different keywords appearing in the LREC proceedings from 2000 to 2014. Then, we compute the tf-idf values of the keywords for each LREC proceedings separately. In this way, we are able to produce a time-series data for each keyword consisting of successive tf-idf measurements over a time period from 2000 to 2014. Table 1 shows some example extracted keywords, their ranking and tf-idf values for different LREC proceedings. Here, we can clearly see some emerging trends for Neural Network and Linked Data especially after 2010. This dataset is our key contribution in this work, and is publicly available for the research community. It also contains ranked lists of related keywords and researchers for each keyword. This could be

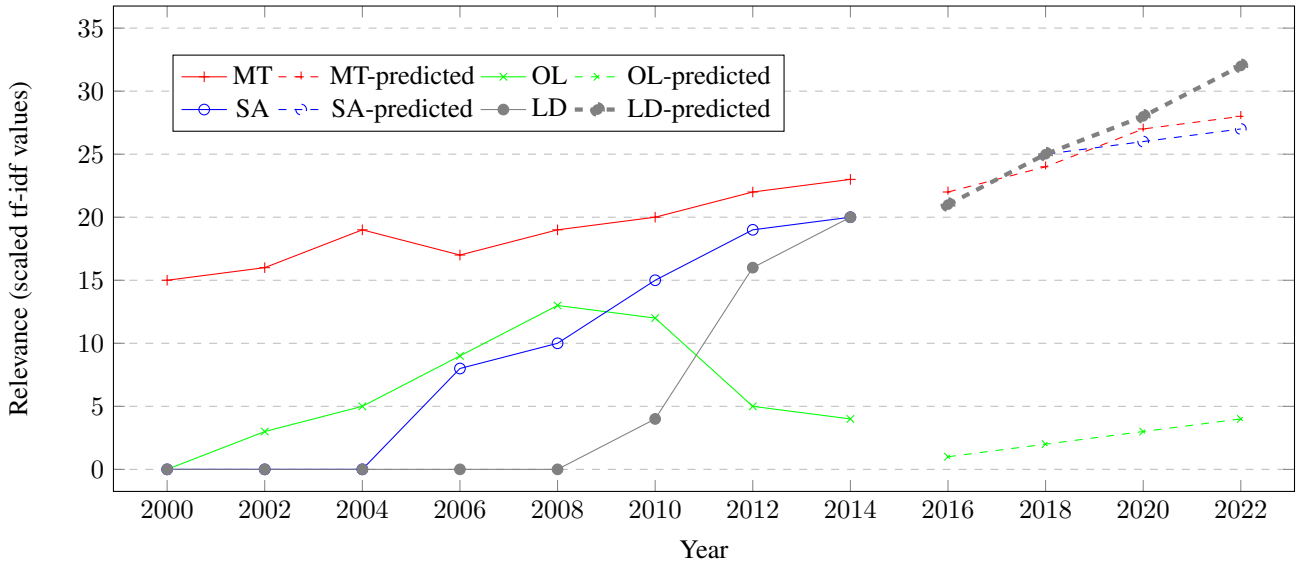

Figure 1: Future forecast for some topics based on Individual Polynomial Models (MT: Machine Translation, OL: Ontology Learning, SA: Sentiment Analysis, LD: Linked Data)

Topic	2014		2012		2010		2008	
	Rank	tf-idf	Rank	tf-idf	Rank	tf-idf	Rank	tf-idf
Natural Language Processing	1	29.12	1	26.87	1	25.99	1	25.64
Language Resources	2	25.97	2	26.05	2	24.52	2	24.23
Sentiment Analysis	13	20.44	16	19.04	53	15.42	288	9.93
Parallel Corpus	29	17.62	39	16.54	46	15.85	51	15.05
Linked Open Data	19	19.49	126	12.76	NA	NA	NA	NA
Neural Network	158	12.30	3646	4.16	8932	2.20	156	11.94
Linked Data	15	19.73	37	16.59	6735	2.77	NA	NA
Linguistic Linked Data	1759	5.83	5357	3.29	NA	NA	NA	NA

Table 1: Sample LREC keyword data: Some keywords with their relevance scores and rankings in 2014, 2012, 2010 and 2008.

really useful as many keywords can be clustered, defining an abstract topic. All this information is extracted using the Saffron framework. Saffron applies a recent approach for keyword extraction based on domain modeling (Bordea et al., 2013), and expert finding and profiling to provide the related experts on different topics.

5.2. Results and Discussion

Apart from the standard forecast error based evaluation, we also compute Spearman correlation against 2014 LREC observation values while learning on the past observations. Table 2 gives a comparison based on Spearman rank correlation and average root mean square (RMS) error for the fine grained models and the global model over the whole data. It is clear that having multiple fine grained models for different keywords capture the temporal growth better than having a single averaged out linear regression model.
Using these models, we can predict the future growth of the topics in LREC. Figure 1 shows a future forecast for some example keywords based on individual polynomial models.

In this work, we simply utilized the keyword's tf-idf value for curve fitting. However, the dataset additionally provides the related topics and experts with relatedness scores with each keyword. These features can also be used as dependent variables, which might produce a better curve fit. Also, more sophisticated approaches for modelling the time series data such as using Fourier transforms or Recurrent Neural Networks can be investigated in context of modelling the growth of scientific topics in this domain. However, we do not have enough temporal data points for LREC (2000-2014) to experiment with such approaches.

Model	Spearman Corr.	RMS Error
Individual Linears	0.74	230.20
Individual Polynomials	0.77	224.42
Global Linear	0.71	254.25
Global Polynomial	0.68	271.74

Table 2: Comparison based on Spearman correlation on LREC 2014 data

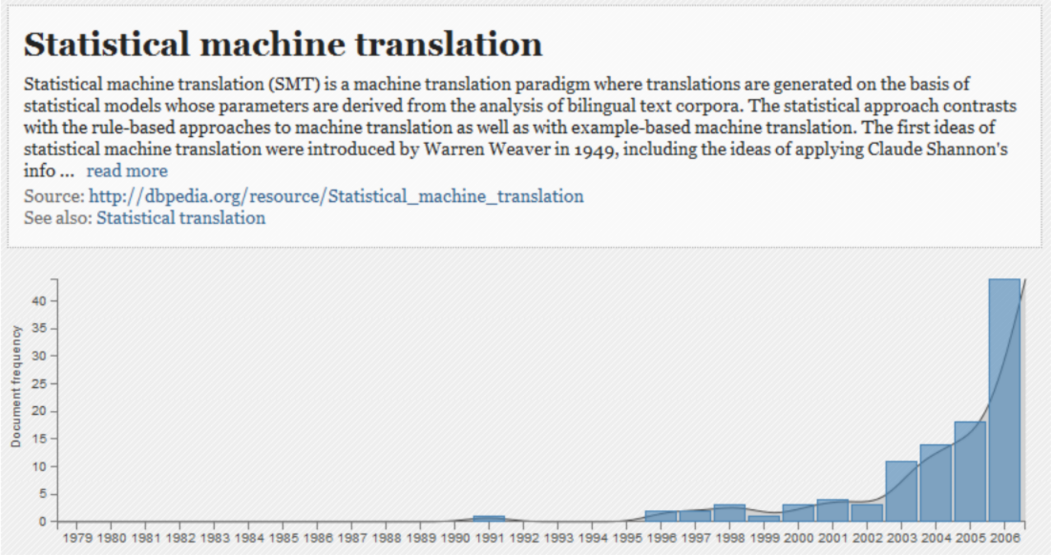

Figure 2: Trend Analysis on "Statistical Machine Translation" in the ACL Anthology

6. Conclusion

In this work, we evaluated a basic approach for forecasting emerging trends in LREC by predicting keyword distribution through regression models. We generated a time series dataset of topics and their popularity based on LREC proceedings. The work presented here can be extended by analyzing additional information from other sources such as publications from other conferences in the same domain, e.g., ACL Anthology, or publication citations (He et al., 2009). We also plan to analyze and evaluate the effect of related topics and authors on forecasting emerging trends. Another interesting direction of extending this work is to apply neural network architectures exhibiting dynamic temporal behaviour such as Recurrent Neural Networks for modelling the time series data of the scientific topics in this domain.

7. Acknowledgements

This work was funded by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (INSIGHT).

8. References

Adar, E. and Datta, S. (2015). Building a scientific concept hierarchy database (schbase). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 606–615.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April.

Bolelli, L., Ertekin, Ş., and Giles, C. L. (2009). Topic and trend detection in text collections using latent dirichlet allocation. In *Advances in Information Retrieval*, pages 776–780. Springer.

Bordea, G., Buitelaar, P., and Polajnar, T. (2013). Domain-independent term extraction through domain modelling. In *the 10th International Conference on Terminology and Artificial Intelligence (TIA 2013), Paris, France*. 10th International Conference on Terminology and Artificial Intelligence.

De Gooijer, J. G. and Hyndman, R. J. (2006). 25 years of time series forecasting. *International journal of forecasting*, 22(3):443–473.

Guo, H., Weingart, S., and Borner, K. (2011). Mixed-indicators model for identifying emerging research areas. *Scientometrics*, 89(1):421–435.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., and Giles, L. (2009). Detecting topic evolution in scientific literature: How can citations help? In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 957–966, New York, NY, USA. ACM.

Lent, B., Agrawal, R., and Srikant, R. (1997). Discovering trends in text databases. In David Heckerman, et al., editors, *KDD*, pages 227–230. AAAI Press.

Mörchen, F., Dejori, M., Fradkin, D., Etienne, J., Wachmann, B., and Bundschus, M. (2008). Anticipating annotations and emerging trends in biomedical literature. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 954–962, New York, NY, USA. ACM.

Rotolo, D., Hicks, D., and Martin, B. (2015). What is an emerging technology? *Research Policy*, 44(10):1827–1843.

Wang, C., Blei, D. M., and Heckerman, D. (2008). Continuous time dynamic topic models. In David A. McAllester et al., editors, *UAI*, pages 579–586. AUAI Press.

Zhou, D., Ji, X., Zha, H., and Giles, C. L. (2006). Topic evolution and social interactions: how authors effect research. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 248–257. ACM.