AIRTA: An Automatic Inter-disciplinary Research Topic Advisor
- Where are We and Where do We Go -

Michael Paul and Roxana Girju
{mjpaul2, girju}@illinois.edu
Linguistics and Computer Science Departments
Beckman Institute
University of Illinois at Urbana-Champaign

No one can predict (at least not in detail) the stringent issues that science and society will consider in the next decades. However, if we look at some top-priority issues of today - such as health, economy, homeland security, stem-cell research, science teaching - and pressing research questions, such as how to enhance child development and learning and even how to make sense of the huge amount of information with which we deal daily, we can say that future topics will be so complex as to require insights from multiple disciplines.

Interdisciplinary research will thus facilitate the integration of information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or sources of specialized knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or field of research.

We believe that like other disciplines, computational linguistics will drastically benefit from an inter-disciplinary perspective since cutting edge research and advancements invariably lie at the boundaries of fields' silos. One of the current projects tackled by our group, Semantic Frontiers at the University of Illinois at Urbana-Champaign deals exactly with this issue. AIRTA - *Automatic Interdisciplinary Research Topic Advisor* - is a tool which is designed to foster interdisciplinary research in order to make breakthrough predictions for future directions. The tool is also designed to promote interdisciplinary collaborations by providing novel topic suggestions to professionals who would like to engage in research discussions with other parties, but who are not familiar with those areas.

AIRTA receives as input two or more disciplines and displays a series of statistics about the evolution of topics in those fields as well as suggestions for novel topics and research trends.

Currently our system displays a series of statistics and graphics which show the dynamics of topics and (local and global) trends in computational linguistics (CL) research based on the proceedings of the major conferences in the field over many years. These statistics give a nice overview of which topics had been highly researched in a period of time as well as past and current novel directions such as question answering and textual entailment.

AIRTA is currently "researching" a set of four fields: linguistics, machine learning, education (we include here educational psychology), and the language technology industry. At the current stage, however, the tool is able to suggest novel topics (those which have never been proposed in the CL community) by combining research ideas from linguistics and machine learning, as well as predicting future trends from this combination.

Our corpus consists of 4,700 papers from the ACL Anthology, 2,300 papers from linguistics journals (Language, Journal of Linguistics, Linguistic Inquiry, Linguistics and Philosophy), and 1,700 papers from education journals (Journal of Education, Educational Psychology). We avoided journals that cover one specific area so that we have a more general representation of each field. Workshops from the ACL Anthology were excluded for the same reason.

The papers from each field were grouped into topics using Latent Dirichlet Allocation (LDA; Blei et al. (2003)) and other methods as appropriate. We ran LDA on the education papers using the words from the titles and abstracts, as the full text was not available. We used LDA to induce 30 topics and chose 18 that seemed relevant. We repeated the process on a subset of these papers that are related to language and grouped these papers into 8 topics. For the ACL Anthology, we loosely followed the methodology of Hall et al. (2008). We ran LDA with 100 topics on the text of the papers and the topics that were relevant.

The linguistics papers were more difficult to label as abstracts were only available from 400 of them, but we had some help in that 340 papers from Linguistic Inquiry were already categorized. We looked at words from the abstracts and titles of these pre-labeled categories and manually reduced these to small lists of discriminative keywords. Weights were assigned to words based on frequency within a category, and these weights were adjusted by hand where necessary. We also manually added new topics, and then assigned papers to 47 topics based on matches from their titles and abstracts to the topic keywords. In each case, papers are assigned to topics by some degree of membership, and there is overlap between topics. The next step is to examine trends in these fields over time.

Trends in linguistics:

There was a sudden increase in phonology and phonetics in the mid-1990s, but these topics have since declined. Pragmatics was fairly prominent in the early 80s, but has remained a minor topic since the mid-80s. Other than a small dip in the early 2000s, the topic of case has increased markedly over the past 20 years. The topic of agreement has gradually but noticeably increased since the early 90s. Discourse has steadily declined since the late 80s.

Trends in education:

Papers in education relating to language and reading have increased dramatically over the past 20 years. The topic of spelling declined noticeably through the 1990s but then increased at the same rate through the 2000s. Papers on bilingualism and foreign languages have increased steadily since the late 90s, although it may be declining again. Reading and literacy is by far the most prominent topic, accounting for about a third of language-related papers in the education field.

The prominence of language acquisition has declined since the mid-late 1990s in both linguistics and education, but is still a fairly prominent topic.

For clues as to why these fluctuate, we look for correlations between changes in gradients of different topics, which might be indicative of how research in some areas affect others. This sort of information will be used to guide our topic advisor. New ideas will be generated in areas that appear to be important.

References

D. Blei, A. Ng, and M. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993-1022.

D. Hall, D. Jurafsky, D. Manning. Studying the History of Ideas Using Topic Models. *EMNLP* 2008.