Scientific Document Analysis
in general and in the medical domain

BioMedical NLP

Lect Dr Ana-Sabina Uban
auban@fmi.unibuc.ro

Scientific documents

Scientific document analysis = *bibliometrics / scientometrics*

Why analyze scientific data?

Better understanding of science and the scientific process (How do researchers choose topics of research? Does double-blind review work? Can we guess the author of a scientific text automatically? Can we review or summarize a scientific article automatically? Are current metrics used for measuring scientific quality good enough?)

Help with scientific discoveries: corroborate and link ideas and results from different scientific articles to produce new knowledge (discovering scientific ideas with word embeddings, ... measuring scientific trends,)

In the medical domain:

Organizing medical information for fast retrieval across sources, enabling and accelerating medical/biology research, combining information for discovering new insights, new drugs, drug interactions etc, optimization of clinical trials.

More data readily available compared to clinical data

Scientific documents

Sources:

Repositories, Journals and Conference Proceedings:

- e.g. ACL conference suite proceedings (most important publications in computational linguistics & NLP):
  - https://aclanthology.org/
- Medical articles repositories (see next slide)
- Search engines and web-based browsers for science repositories:
  - https://scholar.google.com (relevant metadata: publication venue, year of publication, number of citations)
  - https://pubmed.gov (medical articles browser, allows search by metadata such as relevant terms/categories)

Scientific documents

Format:

Depth-resolved analysis of spontaneous phase separation in the growth of lattice-matched AlInN

title

A. Redondo-Cubero[1,2,*], K. Lorenz[3], R. Gago[4], N. Franco[3], M.-A. di Forte Poisson[5], E. Alves[3] and E. Muñoz[1]

authors

1 ISOM and Dpt. de Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, E-28040 Madrid, Spain.

2 Centro de Micro-Análisis de Materiales, Universidad Autónoma de Madrid, E-28049 Madrid, Spain.

3 Instituto Tecnológico e Nuclear, Estrada nacional 10, 2686-953 Sacavém, Portugal.

4 Instituto de Ciencia de Materiales de Madrid (CSIC), E-28049 Madrid, Spain.

5 Thales Research & Technology/TIGER, 91461 Marcoussis Cedex, France.

affiliation

ABSTRACT:

We report the detection of phase separation of an $Al_{1-x}In_xN$/GaN heterojunction grown close to lattice matched conditions (x~0.18) by means of Rutherford backscattering spectrometry in channeling geometry and high resolution x-ray diffraction. An initial pseudomorphic growth of the film was found, with good single crystalline quality, the

abstract

Figure 1: Block segmentation of PDF documents before construing content

Scientific documents

Format:

- *Title, abstract, authors & affiliations, keywords*
- *Introduction, Related work*
- For ML or corpus-based studies: Section on *Dataset*
- *Methodology* (be it algorithms, ML models, or clinical study methodology etc)
- *Results* (be it classification results, clinical study results etc)
- *Discussion, Conclusions*

Scientific documents

Processing of scientific articles:

Scientific article processing tool: **Grobid**: https://github.com/kermitt2/grobid

- parse scientific documents (from input PDFs) as structured text (XML format)
- separately extract title, abstract, authors, headers, citations, references

Scientific medical documents

Corpora of medical scientific literature
and relevant institutions

US National Library of Medicine (NLM)

- 1818: Initially a resource for military physicians (the Surgeon's General Library) since 1818; 2,300 volumes
- 1870s: + 120,000 volumes in the 1870s + introduced Index Medicus (index of the journal articles in the Library).
- 1956: it became the USA national library in the Public Health Service: **NLM** - National Library of Medicine. (John F. Kennedy)
- 1960s: specifications for computerized system for producing the monthly Index Medicus => MEDLARS (Medical Literature Analysis and Retrieval System) - provided individualized bibliographies through specialized "demand search" => keeping up with the demand eventually raised the need for automatic information retrieval

Corpora of medical scientific literature
and relevant institutions

=> NLM issued a policy that stated “the need for a national resource for information systems research and development relevant for human health” and “a clearing-house and coordinating agency for information systems R and D within the Public Health Service”
1988: **NCBI** (National Center for Biotechnology Information) - division of NLM

Link: https://www.nlm.nih.gov/bsd/mmshome.html

Corpora of medical scientific literature
and relevant institutions

MEDLINE database

Medical Literature Analysis and Retrieval System Online

Large bibliographic database maintained by NLM: authoritative and comprehensive source of references into peer reviewed fundamental biomedical studies, and clinical evidence
20 million references to articles from 4.800 biomedical journals in 30 languages
1,500-3,000 references added everyday
Each citation includes: title, authors, journal, type, date etc; + link to full paper from original publisher, many of them free

Link: http://www.nlm.nih.gov/bsd/

Corpora of medical scientific literature
and relevant institutions

PubMed

MEDLINE is accessible on the web through PubMed, NLM's gateway
PubMed is a boolean search engine that indexes titles, abstracts and metadata separately + query augmentation for recognized terms

over 7.5 million full-text biomedical articles

+ **Entrez** Programming Utilities for batch retrieval of MEDLINE citations

Link: https://pubmed.ncbi.nlm.nih.gov/

Corpora of medical scientific literature
and relevant institutions

PubMed Central International

PubMed Central: digital archive of biomedical literature (full texts of articles), among other such archives
Part of PubMed Central International network, which includes US PubMed Central and UK PubMed Central, PubMed Canada - they archive articles from journals and manuscripts funded by NIH and Welcome Trust and Canadian Institutes of Health Research respectively ([where they have rights to the content])
A common format format is used (the Journal Archiving and Interchange Tag Suite), in view of development of the global network for exchange of biomedical information

Link: https://www.ncbi.nlm.nih.gov/labs/pmc/

Corpora of medical scientific literature
and relevant institutions

Medical Subject Headings (MeSH)

Additional metadata for each MEDLINE citation.

MeSH: the thesaurus of NLM controlled vocabulary
23,000 descriptors arranged in a hierarchical structure
151,000 Supplementary Concept Records (additional chemical substance names)
+ software assistive tools since 2002: MTI (Medical Text Indexer) that automatically suggests MeSH headings (Aronson et al 2004)

Link: https://www.nlm.nih.gov/mesh/meshhome.html - downloadable! ; browser: https://meshb.nlm.nih.gov/

Corpora of medical scientific literature
and relevant institutions

GENIA

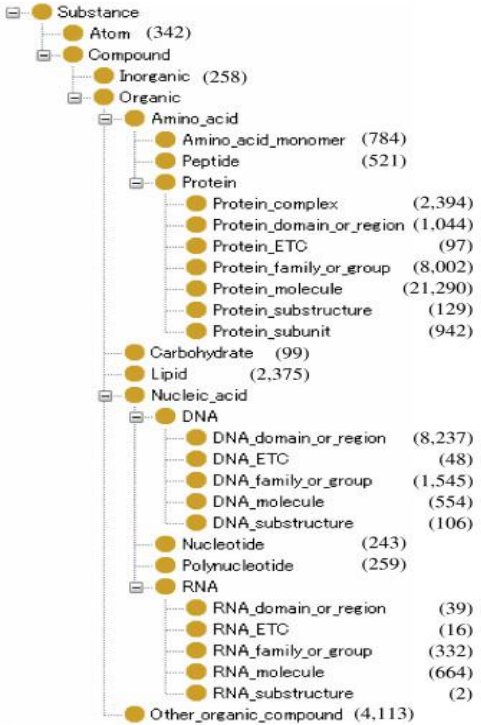
GENome Information Acquisition project: enhancing the MEDLINE knowledge base for specific purposes: automatic extraction of biochemical information from journal papers and abstracts by means of language engineering

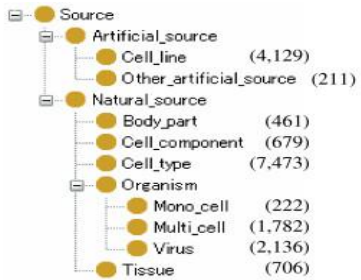
Developed eventually into a set of tools and annotated documents for development and use of BioNLP methods

Corpora of medical scientific literature

and relevant institutions

GENIA Ontology





Corpora of medical scientific literature
and relevant institutions

GENIA

The corpus contains 1,999 MEDLINE selected abstracts, annotated with linguistic and semantic information from GENIA ontology (https://dl.acm.org/doi/pdf/10.5555/1289189.1289260)

Tools: POS tagger, syntactic parser, shallow parsing and NER; annotation tool (XConc) = XML editor, concordancer, and ontology browser; tool for protein-protein interactions extraction
Available for download: http://www.geniaproject.org/home

Information extraction from medical articles

Medical scientific documents

Problems and tasks in medical scientific document processing:

- information extraction and information retrieval in general (useful for downstream applications): NER, relation extraction
- automatic MeSH article indexing & ontology enriching
- drug discovery, drug repurposing, drug-drug interaction, drug-target interaction, (“in silico” literature-based discovery)
- scientific trend monitoring (meta-research; identifying understudied diseases etc; commercial: identify drugs & research opportunities)

Information extraction for medical articles

Datasets (corpora) for information and relation extraction with NLP (useful for supervised learning):

- Shared tasks annotated corpora:
  - 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text - PMC [1]
  - SemEval 2013 on drug-drug interactions (DDIExtraction): [2,3]
- A Dataset for N-ary Relation Extraction of Drug Combinations [4].
- RareDis: a corpus annotated with rare diseases, their signs and symptoms [5]

A recent survey on recent methods for biomedical literature mining [6].

Information extraction for medical articles
Examples

On the **motivation** for using automation for identifying meaningful **drug–target interactions (DTIs) that were not reported in DrugBank**: [7]

- "*Knowledge bases play an important role in the development of biomedical science. Most structured databases, such as DrugBank1, CTD2, SIDER3 and BioGRID4, are curated from a large number of scientific articles by human experts, who expend huge amounts of time and effort. Biomedical information extraction technology aims to shift this time-consuming and tedious burden to machines by developing efficient computational tools to extract meaningful facts from vast unstructured texts automatically. After that, often with some human curation, the extracted data can be fed into the downstream tasks to facilitate the related biological knowledge discovery processes. The information that biomedical researchers most care about generally falls into three types: biomedical entities, relations (interactions or associations between entities) and events (important facts or findings attached to at least one entity)*"

Information extraction for medical articles
Some recent studies

- Neural methods for inter-sentence relation-extraction (where entities are potentially located in different sentences - since relation extraction was mostly done for entities that are in the same sentence in the past): [8]
- Notable references on information/relation extraction using advanced NLP, including most recent approaches: [9,10,11,12]
- Submissions to i2b2 shared tasks on information extraction [13] and DDIExtraction [2,3]

Information extraction for medical articles
Some recent studies - transformer-based

Information/relation extraction for the biomedical domain using **transformer-based prompting**:

- biomedical knowledge probing [14],
- biomedical entity normalization [15],
- biomedical relation extraction [16, 17],
- Biomedical paraphrasing for understanding rare terms [18]

Information extraction for medical articles
Automatic indexers

Indexing medical articles involves in many cases manually-validated metadata for annotating medical publications with relevant tags (diseases mentioned, treatments, compounds, etc). Metadata of every new article on PubMed is manually validated and updated yearly.

There is the question of the place of NLP in complementing this information.

One task is:

Predicting MeSH terms and changes in the MeSH ontology before it is actually done manually (ontology changes appear yearly) using NLP

Information extraction for medical articles
Automatic indexers

- In the recent years, there has been an increased attention from the NLP community in developing **automatic MeSH indexers**.
- Several methods include Graph Convolution Networks [19] and attention-based methods [20, 21],

Information extraction for medical articles
Automatic indexers

Some proposed solutions:

- An approach using information extraction, classification based on topic modelling, clustering/ranking and feature extraction based on trend tracking. [8]. On motivation: "*Currently, PubMed abstracts are annotated mostly manually by human experts who read the entire document before adding MeSH terms. This procedure causes a massive time delay between the inclusion of a new document in PubMed and its annotated version. The MeSH ontology is also curated manually undergoing a major update every year based on newly upcoming technologies or findings in the biomedical field. For example, genes or proteins that are crucial for the development and progress of a certain disease are included as concepts to the MeSH ontology once a certain degree of consensus in the scientific community has been established*".

Information extraction for medical articles
Automatic indexers

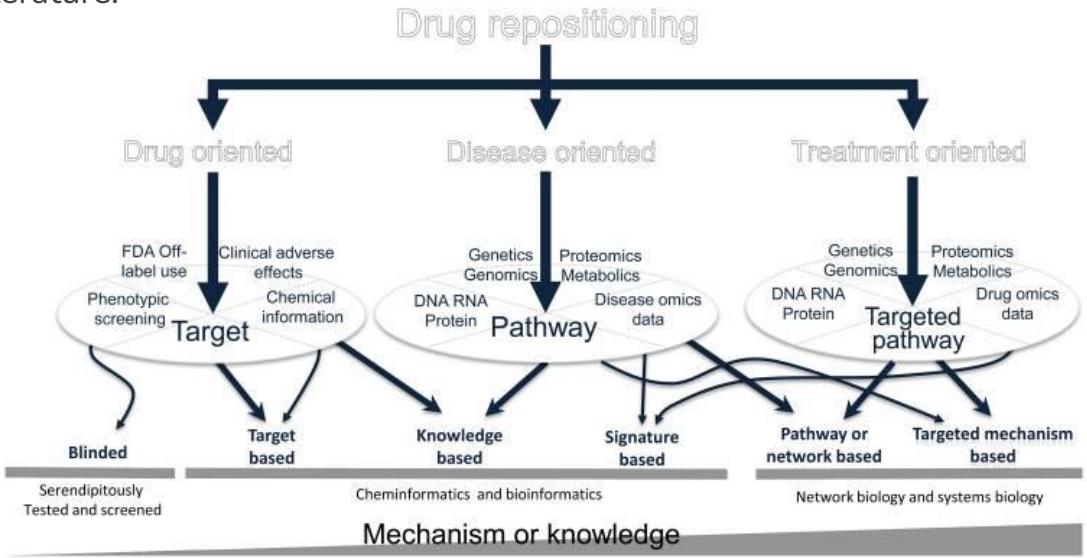
- Recently - **transformer-based methods: BERTMeSH** [22].
- Short review from [7]: "*BioASQ, an EU-funded project, has organized challenges on automatic MeSH indexing since 2013, which provides opportunities to involve more participants in continuing to the development of MeSH indexing systems. Many effective MeSH indexing systems have been developed since then, such as MeSHLabeler (Liu et al., 2015), DeepMeSH (Peng et al., 2016), AttentionMeSH (Jin et al., 2018), and MeSHProbeNet (Xun et al.2019).*"

Information extraction for Drug discovery & drug repurposing

Drug repurposing

Drug repurposing is a strategy for discovering new uses for existing drugs.

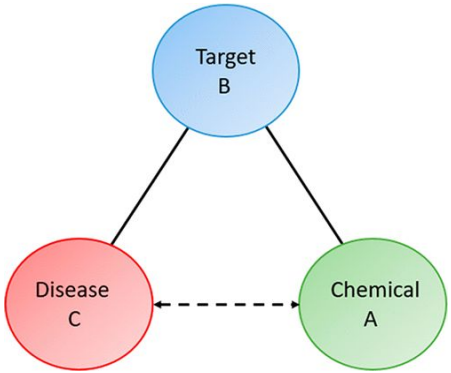
In silico approaches: using computational methods, often based on mining scientific literature.



Drug–Target–Disease Relationships (DTD)

"*Elucidation of the mechanistic relationships between drugs, their targets, and diseases is at the core of modern drug discovery research*"

[23]: The authors mine the biomedical literature to identify relationships between the drugs, biological targets, and diseases (annotated by MeSH terms) that co-occur in the abstracts of PubMed papers. The authors use an extension of Swanson's ABC paradigm. Chemical A is supposed to have an effect on Disease C because both are associated (co-occurs in the same articles) with the Target B.



Drug-target interactions (DTIs)

Extracting information about drug-target information is critical for drug repurposing and elucidation of drug mechanisms.
[9] use a BERT-based method for identifying abstracts containing information about drug-target interactions. *After detecting abstracts with DTIs, drug and protein entities were annotated using the PubTator dataset. Further, the BERT models predicted the assay format (cell-based, organism-based or other assays) used in the documents.*

Drug-disease treatment pairs

Used for drug repurposing in two ways [24]:

- Case studies or small-scale clinical studies already contain evidence for drug-disease repurposing pairs
- Discovering similar drugs to the ones from the drug-disease pairs that can have the potential to treat the disease

[24] use a pattern-learning approach for extracting treatment-specific drug-disease pairs from MEDLINE abstracts.
[10] extract disease-medication associations from clinical data (not scientific) "**Distributional Relation Extraction** (DRE) - a supervised machine learning method for discovering associations between given entity pairs using distributional semantics and UMLS."

Novel drug-disease associations (DDAs)

Network based approaches:

- Novel DDAs can be predicted from heterogeneous networks containing information about drug-drug similarity, disease-disease similarity, and known drug-disease associations. After learning the network embedding vectors, a classifier is trained to predict the probabilities of drugs and diseases being associated [25].

Detecting off-target effects

Most targeted drugs display a wide range of target selectivity, and do not have only a specific target. Unintended off-target effects can help in revealing potentially new cancer vulnerability; thus harnessing these effects can lead to novel and promising precision medicine approaches. [26].

In NLP, the detecting of off-target effects was not used for drug repurposing, but rather for discovering the off-target side effects of drugs [27].

Scientific Trend Detection

With a case study on analyzing evolution of topics in a scientific subcommunity:
NLP & DL conferences

Trend discovery in medical scientific documents

General

A survey from 2004 on detecting emerging scientific trends [32]

Some approaches:

- Citation-based:
  - Temporal document clustering (cluster papers within the same timeframe based on keywords) and Journal co-citation clustering (clustering journals based on co-citations => potential trends) [33]
  - Discovers novelty in recombining existing knowledge based on citation graphs [34]
  - Detecting emerging trends and transient patterns in scientific literature [35]

Trend discovery in medical scientific documents

Citations and time-series-based methods for trend extraction:

- Finding innovative papers based on outlier cited papers [28]
- Statistical methods on paper topics/citations over time [29]
- Prediction of “topic” trends based on MeSH terms: detecting core topics and trending new topics (focused on a selection of MeSH headings). [30]
- Catching trends using MeSH term analysis - based on statistical tests applied on frequency time series [31]

Trend discovery in computational linguistics papers [36]

» Scientific Ideas

Model

Distributional model of ideas based on topic modelling (LDA).

$$\textbf{idea}_i := (\textbf{topic}_t, \textbf{type}_c), \quad (4)$$

$type_c \in \{task, algorithm, data\}$, $topic_t \in T$,
T = the set of topics generated by the LDA algorithm.

Dataset

ACL+: 11,092 articles published in computational linguistics conferences: ACL, EMNLP, COLING, NAACL (1964-2018).
NIPS: 6,893 articles published in NIPS (1994-2018).
100 topics trained on ACL+ corpus; manually curated to 84 topics.

» Prevalent topics

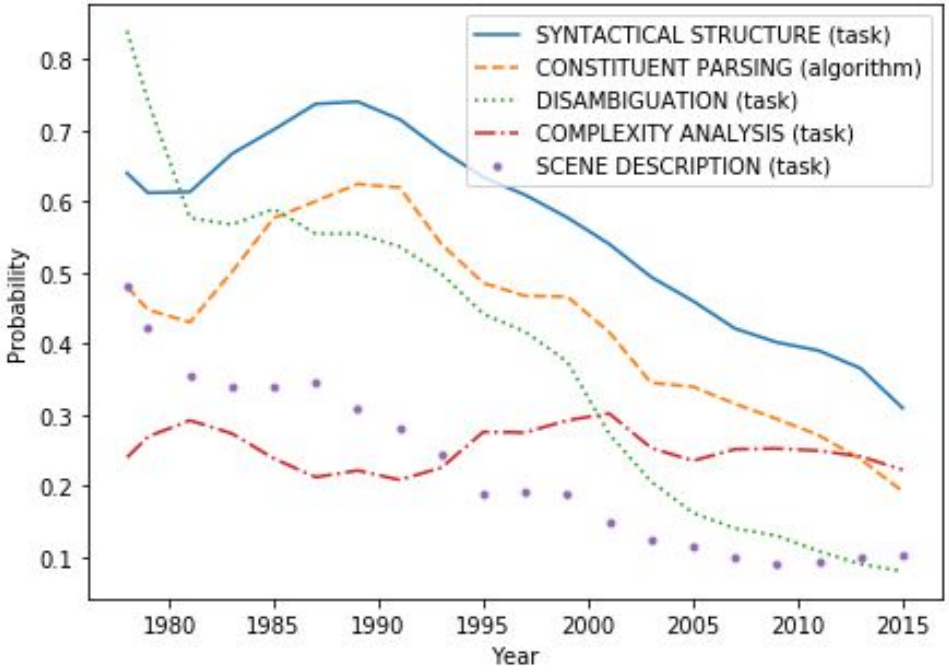


Figure: Most prevalent ideas in ACL+

» Varying topics

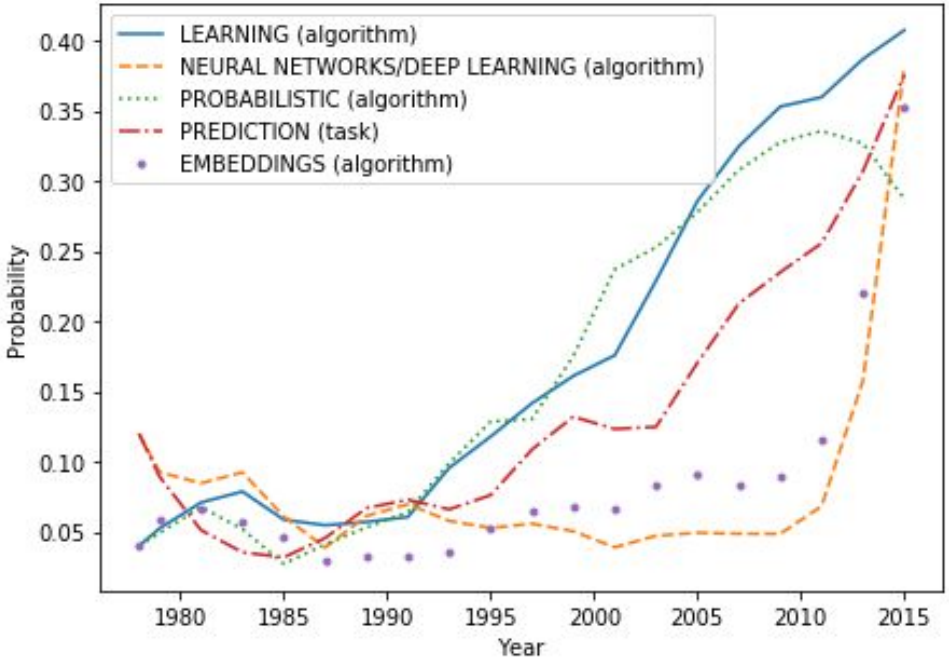


Figure: Top varying topics in ACL+

» Idea Relationships

Types of cooperating and competing relationships between pairs of ideas based on their:

* Co-occurrence in the same documents (PMI score)
* Correlation of prevalence (probability of occurrence of a topic) over time

For each relationship, pairs of ideas considered according to their types:

* same-type ideas for competing relationships (algorithm-algorithm, task-task, ...)
* different type ideas for cooperating relationships (task-algorithm, algorithm-data, ...)

» Idea Relationships

Friendship. Two ideas are "friends" if:

* correlation of prevalence over time > 0
* co-occurrence in documents score > 0

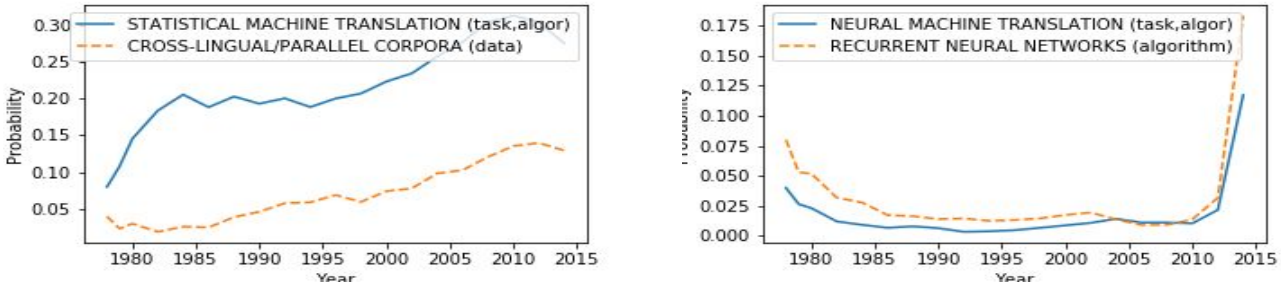


Figure: Friendship relationships.

» Idea Relationships

Head-to-head. Two ideas are "head-to-head" if:

- correlation of prevalence over time < 0
- co-occurrence in documents score < 0

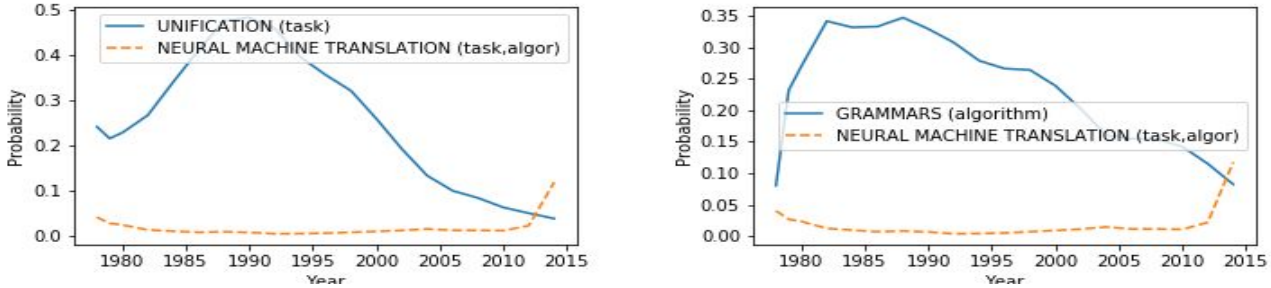


Figure: Head-to-head relationships.

» Idea Relationships

Arms race. Two ideas are at "arms race" if:

- correlation of prevalence over time > 0
- co-occurrence in documents score < 0

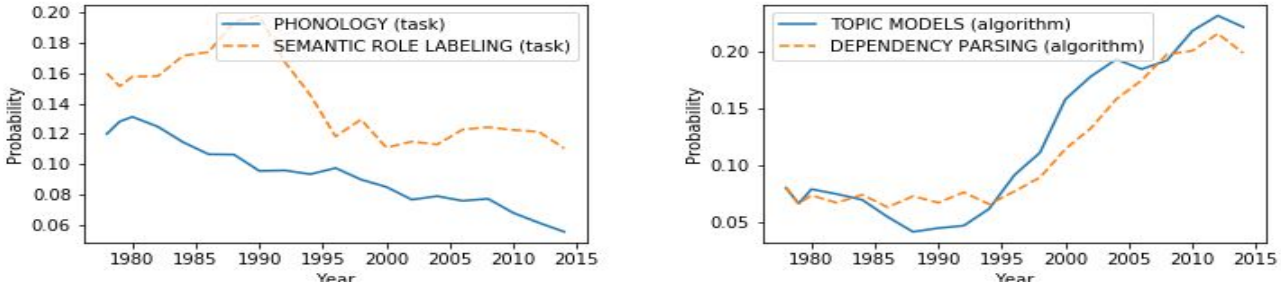


Figure: Arms-race relationships.

» Idea Relationships

Tryst. Two ideas are in a "tryst" relation if:

* correlation of prevalence over time < 0
* co-occurrence in documents score > 0

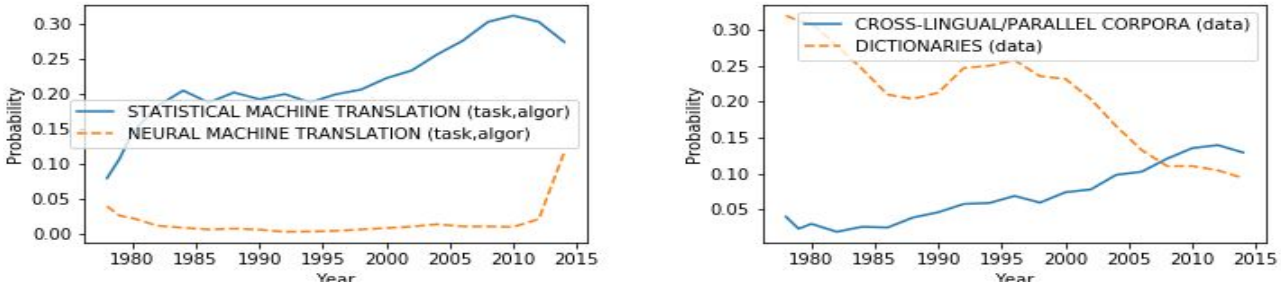


Figure: Tryst relationships.

» Conference Divergence

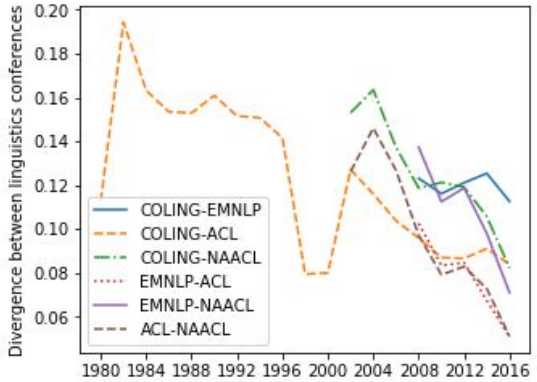
Conference divergence is measured using Janson-Shannon divergence between the topic distributions for articles published in a conference:

$$D_{JS} = \frac{D_{KL}(C_1 || \frac{C_1+C_2}{2}) + D_{KL}(C_2 || \frac{C_1+C_2}{2})}{2} \quad (5)$$

where:
D_{KL} is the Kullback-Leibler divergence
C_i is the average topic distribution for conference conf $_i$:

$$P(C_i) = \sum_{\substack{t \in \\ T}} P(topic = t | conference = conf_i) \quad (6)$$

» Conference divergence



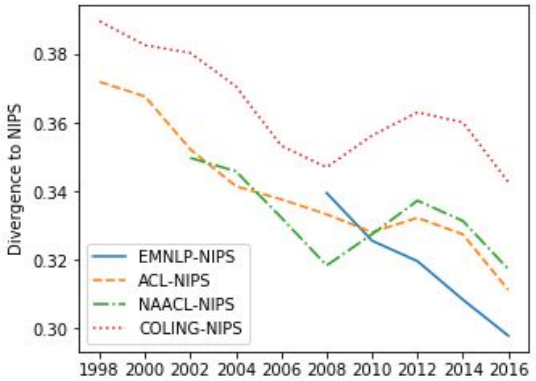


Figure: Conference divergence over time.

» Conference correlation

Conference correlation is computed as the average prevalence correlation across all topics for a given pair of conferences.
Average correlation between ACL+ and NIPS: 0.71

Topic	Correlation
Reinforcement learning	0.93
Finite state machines	0.90
Disambiguation	0.90
Ranking	0.89
Neural machine translation	0.88

Table: Top correlated topics between ACL+ and NIPS.

» Neural networks in ACL+ vs NIPS

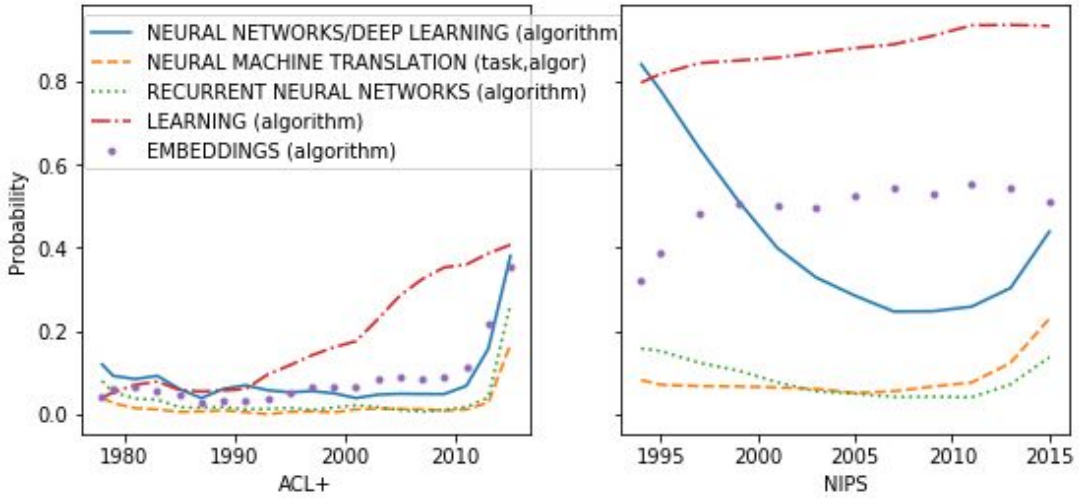


Figure: Neural networks topics in ACL+ vs NIPS.

» Conference synchronicity

Conference synchronicity is a measure of how synchronized two conferences are in terms of topic popularity over time. Computed by measuring conference correlation with an added year lag factor for one of the conferences, at different lag values:

$$P_l(t|y) = P(t|y + l) \tag{7}$$

where l is a lag factor (integer)
and $P(t|y)$ is the probability of topic t occurring in year y.

» Conference synchronicity

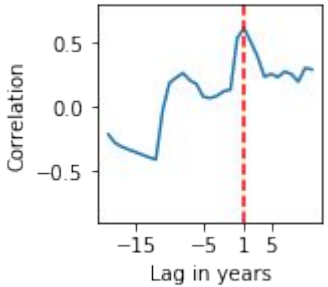


Figure: Neural topics

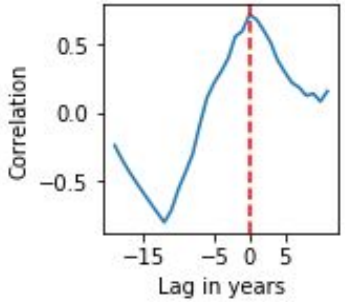


Figure: All topics

Figure: Correlation between topics in ACL+ vs. NIPS with lags.

References

[1] Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):552-6. doi: 10.1136/amiajnl-2011-000203. Epub 2011 Jun 16. PMID: 21685143; PMCID: PMC3168320.

[2] Segura-Bedmar, Isabel, Paloma Martínez Fernández, and María Herrero Zazo. "Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)." Association for Computational Linguistics, 2013.

[3] Segura-Bedmar, Isabel, Paloma Martínez, and María Herrero-Zazo. "Lessons learnt from the DDIExtraction-2013 shared task." Journal of biomedical informatics 51 (2014): 152-164.

[4] Tiktinsky, Aryeh, et al. "A Dataset for N-ary Relation Extraction of Drug Combinations." arXiv preprint arXiv:2205.02289 (2022).

[5] Martínez-deMiguel, Claudia, et al. "The RareDis corpus: a corpus annotated with rare diseases, their signs and symptoms." Journal of Biomedical Informatics 125 (2022): 103961.

[6] Zhao, Sendong, et al. "Recent advances in biomedical literature mining." Briefings in Bioinformatics 22.3 (2021): bbaa057.

[7] Hong, Lixiang, et al. "A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories." Nature Machine Intelligence 2.6 (2020): 347-355.

[8] Noriega-Atala, Enrique, et al. "Neural Architectures for Biological Inter-Sentence Relation Extraction." arXiv preprint arXiv:2112.09288 (2021).

References

[9] Aldahdooh, Jehad, Markus Vähä-Koskela, Jing Tang, and Ziaurrehman Tanoli. "Using BERT to identify drug-target interactions from whole PubMed." BMC bioinformatics 23, no. 1 (2022): 1-13.

[10] Dandala, Bharath, et al. "Scoring disease-medication associations using advanced NLP, machine learning, and multiple content sources." Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016). 2016.

[11] Kavuluru, Ramakanth, and Zhenghao He. "Unsupervised medical subject heading assignment using output label co-occurrence statistics and semantic predications." International Conference on Application of Natural Language to Information Systems. Springer, Berlin, Heidelberg, 2013.

[12] Chun, Hong-Woo, et al. "Extraction of gene-disease relations from Medline using domain dictionaries and machine learning." Biocomputing 2006. 2006. 4-15.

[13] Uzuner, Özlem, et al. "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text." Journal of the American Medical Informatics Association 18.5 (2011): 552-556.

[14] Yao, Zonghai, et al. "Context Variance Evaluation of Pretrained Language Models for Prompt-based Biomedical Knowledge Probing." arXiv preprint arXiv:2211.10265 (2022).

[15] Lai, Zhaohong, et al. "Continuous Prompt Enhanced Biomedical Entity Normalization." Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part II. Cham: Springer Nature Switzerland, 2022.

[16] Yeh, Hui-Syuan, Thomas Lavergne, and Pierre Zweigenbaum. "Decorate the Examples: A Simple Method of Prompt Design for Biomedical Relation Extraction." arXiv preprint arXiv:2204.10360 (2022).

References

[17] Li, Qing, et al. "BioKnowPrompt: Incorporating imprecise knowledge into prompt-tuning verbalizer with biomedical text for relation extraction." Information Sciences 617 (2022): 346-358.

[18] Wang, Haochun, et al. "Prompt Combines Paraphrase: Teaching Pre-trained Models to Understand Rare Biomedical Words." arXiv preprint arXiv:2209.06453 (2022).

[19] Yu, Miaomiao, Yujiu Yang, and Chenhui Li. "HGCN4MeSH: Hybrid Graph Convolution Network for MeSH Indexing." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 20-26. 2020.

[20] Jin, Qiao, Bhuwan Dhingra, William Cohen, and Xinghua Lu. "AttentionMeSH: simple, effective and interpretable automatic MeSH indexer." In Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering, pp. 47-56. 2018.

[21] Wang, Xindi, Robert E. Mercer, and Frank Rudzicz. "KenMeSH: Knowledge-enhanced End-to-end Biomedical Text Labelling." In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2941-2951. 2022.

[22] You, Ronghui, et al. "BERTMeSH: deep contextual representation learning for large-scale high-performance MeSH indexing with full text." Bioinformatics 37.5 (2021): 684-692.

[23] Capuzzi, Stephen J., Thomas E. Thornton, Kammy Liu, Nancy Baker, Wai In Lam, Colin P. O'Banion, Eugene N. Muratov, Diane Pozefsky, and Alexander Tropsha. "Chemotext: a publicly available web server for mining drug–target–disease relationships in PubMed." Journal of chemical information and modeling 58, no. 2 (2018): 212-218.

[24] Xu, Rong, and QuanQiu Wang. "Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing." BMC bioinformatics 14, no. 1 (2013): 1-11.

References

[25] Zhou, Renyi, Zhangli Lu, Huimin Luo, Ju Xiang, Min Zeng, and Min Li. "NEDD: a network embedding based method for predicting drug-disease associations." BMC bioinformatics 21, no. 13 (2020): 1-12.

[26] Palve, Vinayak, Yi Liao, Lily L. Remsing Rix, and Uwe Rix. "Turning liabilities into opportunities: Off-target based drug repurposing in cancer." In Seminars in Cancer Biology, vol. 68, pp. 209-229. Academic Press, 2021.

[27] Zhou, Kaiyin, Sheng Zhang, Xiangyu Meng, Qi Luo, Yuxing Wang, Ke Ding, Yukun Feng, Mo Chen, K. Bretonnel Cohen, and Jingbo Xia. "CRF-LSTM text mining method unveiling the pharmacological mechanism of off-target side effect of anti-multiple myeloma drugs." In Proceedings of the BioNLP 2018 workshop, pp. 166-171. 2018.

[28] King, Daniel, Doug Downey, and Daniel S. Weld. "High-precision extraction of emerging concepts from scientific literature." Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020.

[29] Decker, Sheron Levar, et al. "Detection of bursty and emerging trends towards identification of researchers at the early stage of trends." (2007).

[30] Wang, Liqin, et al. "Content and trends in medical informatics publications over the past two decades." MEDINFO 2017: Precision Healthcare through Informatics. IOS Press, 2017. 968-972.

[31] Ilgisonis, Ekaterina V., et al. "How to catch trends using MeSH terms analysis?." Scientometrics 127.4 (2022): 1953-1967.

[32] Kontostathis, April, et al. "A survey of emerging trend detection in textual data mining." Survey of text mining (2004): 185-224.

References

[33] Tho, Quan Thanh, Siu Cheung Hui, and Alvis Fong. "Web mining for identifying research trends." International conference on Asian digital libraries. Springer, Berlin, Heidelberg, 2003.

[34] Mukherjee, Satyam, et al. "A new method for identifying recombinations of existing knowledge associated with high-impact innovation." Journal of Product Innovation Management 33.2 (2016): 224-236.

[35] Chen, Chaomei. "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature." Journal of the American Society for information Science and Technology 57.3 (2006): 359-377.

[36] Uban, Ana Sabina, Cornelia Caragea, and Liviu P. Dinu. "Studying the Evolution of Scientific Topics and their Relationships." Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021.