EMNLP versus ACL: Analyzing NLP Research Over Time

Sujatha Das Gollapalli, Xiao-Li Li
Institute for Infocomm Research, A*STAR, Singapore
{gollapallis,xlli}@i2r.a-star.edu.sg

Abstract

The conferences ACL (Association for Computational Linguistics) and EMNLP (Empirical Methods in Natural Language Processing) rank among the premier venues that track the research developments in Natural Language Processing and Computational Linguistics. In this paper, we present a study on the research papers of approximately two decades from these two NLP conferences. We apply keyphrase extraction and corpus analysis tools to the proceedings from these venues and propose probabilistic and vector-based representations to represent the topics published in a venue for a given year. Next, similarity metrics are studied over pairs of venue representations to capture the progress of the two venues with respect to each other and over time.

1 Introduction

Scientific findings in a subject-area are typically published in conferences, journals, patents, and books in that domain. These research documents constitute valuable resources from the perspective of data mining applications. For instance, the citation links among research documents are used in computing bibliometric quantities for authors (Alonso et al., 2009) whereas topic models on research corpora are used to distinguish between influential and impactful researchers (Kataria et al., 2011) and to capture temporal topic trends (He et al., 2009).

Despite several potential benefits mentioned above and the free availability of most research proceedings in NLP through the ACL Anthology[1], the topical and temporal aspects of this corpus are yet to be fully studied in current literature. In this paper, we present our study on research proceedings of approximately two decades from two leading NLP conferences, namely ACL and EMNLP, to complement a previous study on this topic by Hall et al (2008). To the best of our knowledge, we are the first to characterize the developments in the NLP domain using a comparative study of two of its leading publication venues. Our contributions are summarized below:

1. We represent the NLP research corpus from approximately two decades as a keyphrase-document matrix and apply Latent Dirichlet Allocation (Blei et al., 2003) to extract coherent topics from it (Newman et al., 2010).

2. We propose two novel representations for summarizing the venue proceedings in a given year. (1) The **probabilistic** representation expresses each venue as a probability distribution over topics, whereas (2) the **TP-ICP** representation captures topics that are the major focus in the venue for a particular year via *Topic Proportion* (TP) as well as topic importance as measured with *inverse corpus proportion* (ICP).

3. We apply Jensen-Shannon divergence and cosine similarity on our proposed venue representations to analyze the venues over time. Specifically, we ask the following questions: What are the popular topics in ACL and EMNLP in a particular year? Is the topical focus in EMNLP different from ACL? How

[1] https://aclweb.org/anthology/

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2002–2006,
Lisbon, Portugal, 17-21 September 2015.

did the topical focus in each venue change over time?

Organization: We describe our novel venue representations and the measures used to compare them in Section 2. The details of our datasets and experiments are presented in Section 3 along with results and observations. We summarize related research in Section 4 before concluding the paper in Section 5.

2 Methods

Let $Y = \{y_1, y_2 \dots y_T\}$ be the consecutive years in which the research proceedings are available from V, set of publication venues under consideration ($V = \{$"ACL", "EMNLP"$\}$ in this paper). If D is the set of all documents over the years, each document $d \in D$ is associated with $\{K_d, y, v\}$ where K_d refers to the content of d whereas v and y refer to the venue and year in which d was published.

2.1 Venues as Probability Distributions

Let $t_1, t_2 \dots t_k$ denote the topics capturing the content of documents in D. Using probabilistic topic modeling and dimension reduction tools such as Latent Dirichlet Allocation or pLSA (Hofmann, 1999; Blei et al., 2003), we extract for each $d \in D$, $P(t_i|d), i = 1 \dots k$, the multinomial distribution over the topics associated with d.

The venue-topic probability distribution $P(t_i|v_y)$ for a given (venue, year) pair $(v = l, y = m)$ can be computed using $D_{l,m}$, the set of documents published in venue l, in the year m. That is,

$$P_{l,m}(t_i) = \frac{1}{|D_{l,m}|} \sum_{d \in D_{l,m}} P(t_i|d) \qquad (1)$$

Note that the above probabilistic representation facilitates a quantitative measure to compare two venues: the divergence between the probability distributions of the two venues. The Kullback−Leibler divergence (KLD) between two (discrete) probability distributions P and Q is given by: $D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$. Due to the unsymmetric nature of KLD, we use the Jensen-Shannon divergence, a symmetric and finite measure ($0 \leq JSD(P||Q) \leq 1$) based on KLD. Let $M = \frac{1}{2}(P + Q)$,

$$JSD(P||Q) = \frac{1}{2}[D_{KL}(P||M) + D_{KL}(Q||M)]$$

2.2 Venues as TP-ICP Vectors

Discrete probability distributions are often represented in computations as normalized vectors. For instance, the $P(t_i|v)$ values comprise the components of a k-dimensional vector. This topic proportion (TP) vector is similar to the normalized term frequency vector commonly used in Information Retrieval (IR) (Manning et al., 2008). TP values are fractions indicating the percentage of a given topic among all topics covered in a particular year. Thus these values are higher for topics that are the major focus in the venue for a particular year .

We also extend *inverse document frequency*, a popular concept that is used to weigh terms in IR (Manning et al., 2008) to describe **Inverse Corpus Proportion** or ICP. Our objective in defining ICP is to capture the importance of a topic by diminishing the effect of topics that are common across all years. Let $\mathbf{TP}_{v,y}(i)$ represents the proportion of topic t_i in venue v for year y, then $ICP(t_i) =$

$$\log\left(\frac{\sum_{y=1}^{|Y|}\sum_{v=1}^{|V|}\sum_{j=1}^{k}\mathbf{TP}_{v,y}(j)}{\sum_{y=1}^{|Y|}\sum_{v=1}^{|V|}\mathbf{TP}_{v,y}(i)}\right) = \left(\frac{|D|}{\sum_{y=1}^{|Y|}\sum_{v=1}^{|V|}\mathbf{TP}_{v,y}(i)}\right)$$

since $\sum_{j=1}^{k}\mathbf{TP}(j) = 1$, $\mathbf{TP}$ being a probability distribution vector and $|Y| \times |V| = |D|$. The TP-ICP vector for a venue is defined as: $[TP(1) \times ICP(1), \dots TP(k) \times ICP(k)]$ and captures in each component the weighted proportion of a topic in that venue for a year. This novel representation can be considered the topic-level counterpart of the popular TF-IDF representation in IR. Given two TP-ICP vectors $P = [p_1, p_2, \dots p_k]$ and $Q = [q_1, q_2, \dots q_k]$, the similarity between them using the cosine measure is given by:

$$cosine(P, Q) = \frac{\sum_{i=1}^{k} p_i.q_i}{||P||_2.||Q||_2}$$

2.3 Keyphrases for representing documents

Corpus analysis tools often use bag-of-words models and term frequencies for representing documents (Heinrich, 2005). However, research documents are often well-structured, and contain various sections with author information, citations,

Topic ID	Top Words
0	System, Dialogue, Dialogue System, Information, Speech Recognition, Speech, Dialogue Manager, Data Collection, User Utterances
1	Model, Training Data, Language Model, Model Parameters, Models, Generative Model, Probabilistic Model
2	Noun Phrases, Other Hand, Head Noun, Future Work, Corpus, Sentence, Language, Method, Japanese Language, Syntactic Structure
3	Evaluation, Evaluation Metrics, Automatic Evaluation, Machine Translation, Human Judgments, Translation Quality
4	Error, Error Correction, Error Types, Spelling Errors, Category, Errors, Lexical Category, Error Rate, Ccg Parser
5	Sentence, First Sentence, Summarization, Sentence Length, Summarization Task, Document Summarization Text Summarization
6	Algorithm, Search Space, Objective Function, Function, Search Algorithm, Model, Optimization Problem, Large Number
7	Language, Information, Sentence, System, Results, Corpus, Approach, Research, Learning, Language Processing, Systems, Machine
8	Rules, Parse Tree, Grammar, Tree Structure, Root Node, Parse Trees, Grammar Rules, Rule Extraction, Rule Set, Elementary Tree
9	Natural Language, System, Language Generation, Information, Generation System, Language, Sentence, System Architecture
10	Sentiment Analysis, Sentiment, Event, Discourse, Sentiment Classification, Discourse Structure, Discourse Relations
11	Dependency Parsing, Dependency, Parsing, Parser, Dependency Tree, Parse Tree, Dependency Parser, Parsing Model, Dependency Trees
12	Words, Other Hand, Natural Language, Other Words, Corpus, Language Processing, Model, Information, TestSet, Language
13	Pos Tagging, Pos Tags, Word Segmentation, Pos Tag, Words, Model, Word Boundaries, Unknown Words, Chinese Word, Pos Tagger
14	Training Data, Training Set, Test Data, Training, Test Set, Data Sets, Data Set, Labeled Data, Training Examples, Unlabeled Data
15	Language, Target Language, Source Language, Machine Translation, Translation, Different Languages, Language Pairs
16	Features, Feature Set, Training Data, Feature Vector, Training Set, Lexical Features, Model, Feature Space, Test Data
17	Clustering Algorithm, Annotation, Same Cluster, Clustering, Clustering Method, Clustering Methods, Annotation Scheme
18	Query, Information Retrieval, Search Engine, Web Search, Search, Search Results, Information, Query Terms, Search Engines
19	Relation, Relation Extraction, Relations, Information Extraction, Semantic Relations, Relation Types, Semantic Relation
20	Topic, Topic Model, Topic Models, Topic Distribution, Same Topic, Topics, Model, Distribution, Topic Modeling, Word Distribution
21	Coreference Resolution, Entity, Same Entity, Coreference, Resolution System, Pronoun Resolution, Anaphora Resolution, Entity Type
22	Machine Translation, Translation, Language Model, Word Alignment, Translation Model, Target Language, Model, Translation Quality
23	Word Sense, Sense Disambiguation, Sense, Word Senses, Words, Different Senses, Target Word, Semantic Relations, Lexical Resources
24	Question, Question Answering, Answer, Questions, Correct Answer, Question Types, System, Textual Entailment, Answer Type
25	Semantic Role, Semantic Roles, Semantic Information, Syntactic Structure, Syntactic Information, Semantic, Parse Tree
26	Machine Learning, Learning, Classification Task, Features, Supervised Learning, Text Classification, Learning Algorithms, Feature
27	Semantic Similarity, Vector Space, Similarity Measure, Word Vectors, Vector, Similarity Measures, Similarity, Semantic Space
28	Language Model, Speech Recognition, Language Models, Word Error, Language, Model, Automatic Speech, Speech
29	Grammar, Language, Natural Language, Lexical Entries, Feature Structure, Feature Structures, Finite Set, Other Hand, Lexical Items
Topics ranked by *Inverse Corpus Proportions*: 3 7 4 17 24 20 21 27 9 28 0 19 6 10 23 2 5 18 15 25 26 29 8 13 14 11 16 1 22 12 Maximum ICP: 4.3533, Minimum ICP: 2.1809, Average ICP: 3.5591	

Table 1: The top words for each topic are shown here after modeling the ACL+EMNLP publications over the years with #topics=30. The topics ranked by their ICP values are shown in the last row to illustrate that ICP values indeed capture the specificity of a topic across the years.

and content-related sections such as *abstract*, *related work*, and *experiments*. To best represent the topical content of these documents, we harness the latest work on keyphrase extraction for research documents and represent documents using keyphrases (Hasan and Ng, 2014).

We use the ExpandRank algorithm (Wan and Xiao, 2008) to extract top n-grams $\forall d \in D$. ExpandRank effectively combines PageRank values on word graphs with text similarity scores between documents to score n-grams for a document and was shown to outperform other unsupervised keyphrase extraction methods on research documents in absence of other information such as citations (Gollapalli and Caragea, 2014).

3 Experiments

Datasets and setup: We crawled the ACLWeb for research papers from EMNLP and ACL from the year 1996 through 2014[2] using the Java-based crawler, Heritrix[3]. The text from the PDF documents was extracted using the PDFBox software[4] after which simple rules similar to the ones used in CiteSeer (Li et al., 2006) were employed to extract the "body" of the research document[5]. The numbers of papers for each year at the end of this process are listed in Table 2. From these numbers, it appears that the paper "intake" in each conference has gone up overall during the last decade although occasionally the increase is due to co-location with related conferences such as IJCNLP and HLT[6].

We construct the keyphrase-document matrix using top-100 keyphrases of each document extracted with ExpandRank. The LDA implementation provided in Mallet (McCallum, 2002) was used to extract topics from this matrix. The LDA algorithm was run along with hyperparameter optimization (Minka, 2003) for different numbers of topics between $10 \ldots 100$ in increments of 10. We use the average corpus likelihood over ten randomly-initialized runs to choose the optimal number of topics that best "explain" the corpus (Heinrich, 2005). As indicated by the left plot in Figure 1 this optimum is obtained when the number of topics is 30.

3.1 Results and Observations

The top phrases that reflect the "theme" captured by a topic are shown in Table 1. As indicated in this table, we are able to extract coherent topics from the corpus using LDA on a document-keyphrase matrix (AlSumait et al., 2009; Newman et al., 2010).

Top research topics in NLP: We select five timepoints $\{1996, 2000, 2005, 2010, 2014\}$ by splitting the 1996-2014 period into roughly-

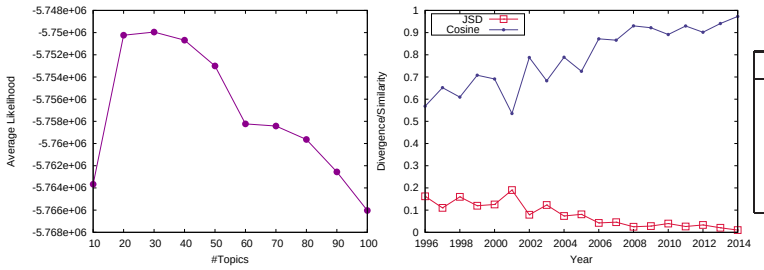
[2] Since our goal is to compare the two venues, we start from 1996 when EMNLP branched off into a full conference from a workshop on Very Large Corpora although ACL proceedings are available from 1979.

[3] https://webarchive.jira.com/wiki/display/Heritrix/Heritrix

[4] https://pdfbox.apache.org/

[5] Processed data available upon request.

[6] ACL was co-located with related conferences in the years 1997, 1998, 2006, 2008, and 2009 and EMNLP in the years 2005, 2007, and 2012.



Year	ACL	EMNLP
2014	27, 10, 20	27, 10, 19
2010	0, 5, 6	20, 19, 18
2005	9, 0, 19	24, 18, 0
2000	7, 9, 5	18, 23, 20
1996	9, 2, 7	28, 2, 5

Figure 1: *Left*: #topics vs. Average Corpus Likelihood, *Middle*: EMNLP vs. ACL, *Right*: Top topics in EMNLP and ACL

Venue	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996
ACL	330	399	227	349	272	244	213	207	310	137	129	103	160	65	45	83	244	73	58
EMNLP	226	207	141	149	125	164	115	131	73	28	53	27	33	10	27	34	13	23	14

Table 2: Number of papers for each venue for different years

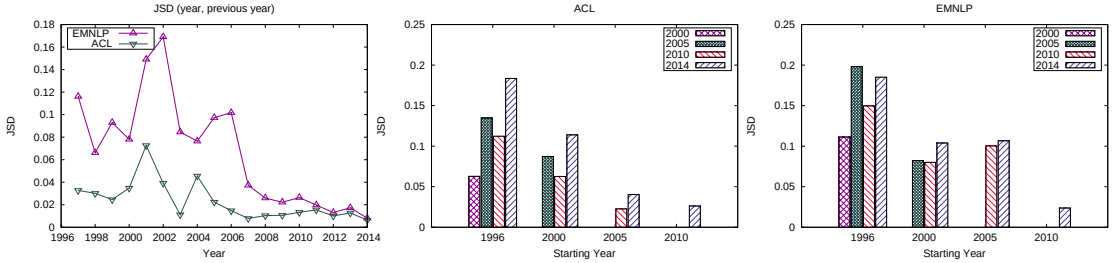


Figure 2: Comparing EMNLP and ACL over the years. Each point in the *Left* plot shows the JSD between a given year y and the year $y-1$. The *Middle* (ACL) and *Right* (EMNLP) plots show the JSD between a timepoint with preceding timepoints in the set $\{1996, 2000, 2005, 2010, 2014\}$.

equal parts and examine the top topics for ACL and EMNLP at these time points. We rank the topics in each conference by their TP-ICP values and list the top 3 topics in the right table of Figure 1. "Semantic relation extraction", "sentiment analysis", and "topic models" are the top research topics in NLP last year (2014) whereas in the year 1996, the topics "noun phrase extraction", "summarization", "corpus modeling", and "speech recognition" dominated the NLP research arena. From the table, it can be seen that "information retrieval" (topicID: 18) ranks among the top topics in EMNLP for all three timepoints during 2000-2010 whereas "natural language generation" (topicID: 9) was consistently addressed during 1996-2005 in ACL.

EMNLP versus ACL: We compare the venues using JSD and Cosine similarity measures in the middle plot of Figure 1. The plot shows decreasing divergence between the topical distributions over the years and increasing cosine similarity between the TP-ICP vectors for the venues. These trends indicate that over the two decades the two venues ACL and EMNLP seem to have "become like each other" although their topical focus was different during the initial years. Increasingly, both venues seem to publish papers on similar topics. This behavior could be interpreted to mean that the NLP research field is more stable now with two of its leading conferences addressing problems on similar topics.

Changing topical focus over the years: In the first plot of Figure 2, we show the Jensen-Shannon divergence between the topic distributions of a particular venue for a given year y and $(y-1)$, the year preceding it. The curve indicates that in the years between 1997-2008, the rate of change from year to year is higher than in the years following 2008. We split the time period 1996-2014 into five roughly-equal parts to form the set $\{1996, 2000, 2005, 2010, 2014\}$. The JSD between the distribution in a particular year and the years preceding it in the above set is shown for ACL (middle plot) and EMNLP (right plot) in Figure 2. For example, the first cluster in the middle plot, shows the JSD values between the distributions for the years 2000, 2005, 2010, 2014 with the starting year 1996 for ACL. For both venues, the divergences of a given year are higher with the early starting years 1996 and 2000 than with the later starting years 2005 and 2010, indicating that the topics being addressed currently in NLP research are significantly different from those addressed a decade back.

4 Related Work

Temporal analysis of corpora is an upcoming research topic in text mining groups. Topic models were particularly investigated for detecting activity patterns in corpora annotated with time information (Huynh et al., 2008; Shen et al., 2009). Evolution of topics and their trends were studied on research corpora from NIPS (Wang and McCallum, 2006) as well as CiteSeer (He et al., 2009).

In contrast with existing approaches that seek to model the detection of new topics and their evolution, we focus on representing different venues pertaining to a research field and examine their development over time by comparing them against each other. In a similar study, Hall et al. (2008) examined the emergence of topics in NLP literature. They proposed "topic entropy" to measure the diversity in three conferences from the ACL Anthology during the years 1978-2006. They also noted that all the venues seem to converge in the topics they cover over the years based on the JSD between their topic distributions.

5 Conclusions

We presented our study on research proceedings of approximately two decades from the leading NLP conference venues: EMNLP and ACL. We extracted coherent topics from this corpus by applying topic modeling on the corresponding keyphrase-document matrix. Next, the extracted topics were used to characterize each venue through probabilistic and vector representations and compared against each other and over the years using various similarity measures. To the best of our knowledge, we are the first to present insights related to the development of a research field by studying two leading conferences in the area using various techniques from NLP and IR.

References

S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, and F. Herrera. 2009. h-index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, 3(4):273 – 289.

Loulwah AlSumait, Daniel Barbar, James Gentle, and Carlotta Domeniconi. 2009. Topic significance ranking of lda generative models. In *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, pages 993–1022.

Sujatha Das Gollapalli and Cornelia Caragea. 2014. Extracting keyphrases from research papers using citation networks. In *AAAI*.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *EMNLP*.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. *ACL*.

Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and C. Lee Giles. 2009. Detecting topic evolution in scientific literature: how can citations help? In *CIKM*, pages 957–966.

G. Heinrich. 2005. Parameter estimation for text analysis. Technical report.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57.

Tâm Huynh, Mario Fritz, and Bernt Schiele. 2008. Discovery of activity patterns using topic models. In *International Conference on Ubiquitous Computing*.

Saurabh Kataria, Prasenjit Mitra, Cornelia Caragea, and C. Lee Giles. 2011. Context sensitive topic models for author influence in document networks. In *IJCAI*, pages 2274–2280.

Huajing Li, Isaac G. Councill, Levent Bolelli, Ding Zhou, Yang Song, Wang-Chien Lee, Anand Sivasubramaniam, and C. Lee Giles. 2006. Citeseerx: a scalable autonomous scientific digital library. In *InfoScale*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Thomas P. Minka. 2003. Estimating a dirichlet distribution. Technical report, Microsoft Research.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies*.

Zhiyong Shen, Ping Luo, Yuhong Xiong, Jun Sun, and Yidong Shen. 2009. Topic modeling for sequences of temporal activities. In *ICDM*, pages 980–985.

Xiaojun Wan and Jianguo Xiao. 2008. Single document keyphrase extraction using neighborhood knowledge. In *AAAI*.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: A non-markov continuous-time model of topical trends. In *SIGKDD*.