Extracting the Evolutionary Backbone of Scientific Domains: The Semantic Main Path Network Approach Based on Citation Context Analysis

Xiaorui Jiang[1], Junjun Liu[2]

[1] Coventry University, Coventry, UK, xiaorui.jiang@coventry.ac.uk

[2] ych1776@163.com

Abstract

Main path analysis is a popular method for extracting the scientific backbone from the citation network of a research domain. Existing approaches ignored the semantic relationships between the citing and cited publications, resulting in several adverse issues, in terms of coherence of main paths and coverage of significant studies. This paper advocated the semantic main path analysis approach to alleviate these issues based on citation function analysis. A wide variety of SciBERT-based deep learning models were designed for identifying citation functions. Semantic citation networks were built by either including important citations, e.g., extension, motivation, usage and similarity, or excluding incidental citations like background and future work. Semantic main path network was built by merging the top-*K* main paths extracted from various time slices of semantic citation network. In addition, this study proposed a three-way framework for quantitative evaluation of main path analysis results. Both qualitative and quantitative analysis on three research areas of computational linguistics demonstrated that, compared to semantics-agnostic counterparts, different types of semantic main path networks provide complementary views scientific knowledge flows. Combining them together, we can obtain a more precise and comprehensive picture of domain evolution and uncover more coherent development pathways between scientific ideas.

Keywords

Main path analysis; semantic main path network; citation function classification; scientific backbone; quantitative main path analysis evaluation

1. Introduction

There were many methods to extract the evolutionary pathways between scientific ideas based on citation network analysis, such as algorithmic historiography (Garfield et al., 2003) and scientific historiograms (Lucio-Arias & Leydesdorff, 2008). Recently, main path analysis (MPA), originally proposed in Hummon and Dorerian (1989), has become popular for extracting the major knowledge diffusion paths among the main ideas advancing an analysed scientific domain, since Batagelj (2003) proposed the efficient *search path counting* algorithms to weight citation edges and Verspagen (2007) laid out the algorithmic foundations for main path extraction.

Most MPA methods were citation semantics-agnostic, i.e., ignoring the semantic relationships between publications. A direct consequence is semantically *incoherent* main path. Figure 1 illustrates a potential cause of this problem – inappropriate search path counts (SPC). In the top-right schematic image, the citation edges (A, B) and (B, C) are both background citations ("Neutral") while the citation edge (A, C) is an extension citation ("Extends"). Ignoring citation function, we have SPC(A, B) ≥ SPC(A, C) because the former is the sum of the number of paths through A→B→C, which is equal to SPC(A, C), and the number of paths through A→B→X(≠C). So traditional MPA approaches will select (A, B), but it is more reasonable to include the extension citation (A, C). Some studies adjusted citation weight by, for example, considering citation preferences according to discipline and publication time (Yu & Pan, 2021) or scaling search path count using citing publication's prestige (Yu & Sheng, 2021). However, the problem was not solved. For example, if B is highly cited, then Yu and Pan's approach will still choose (A, B) in main path exploration. Some weighing schemes used measures of similarity between the abstracts of citing and cited publications (Liu et al., 2014; Huang et al., 2022, Chen et al., 2022). However, such (indirectly inferred) similarity measures shall be less descriptive than authors' own (directly stated) rationales to cite, aka *citation function* (Iqbal et al., 2021; Lyu et al., 2021; Kunnath et al., 2022).

Theoretically, traditional MPA approaches also tend to prefer long local paths[1]. Figure 1 illustrates this case. The left-most image shows a vanilla (semantics-agnostic) main path network (MPN). The longest local path from A00-2018 to D07-1096 is very stretched: `distance`(A00-2018, D07-1096) = 16. It is questionable whether knowledge indeed flows along such long paths with many unimportant citations such as "Neutral". The middle image shows a snapshot of the semantic main path network (semantic MPN) extracted by considering extension ("Ext") and motivation ("Mot") citations. The path becomes more

[1] Refer to Kuan (2022) for more discussions.

compact: `distance`(A00-2018, D07-1096) is decreased to 5. For another example, by further considering usage ("Use") and similarity ("Sim") citations, the longest distance from W96-0213 to W05-0516 is reduced from 17 to 5.

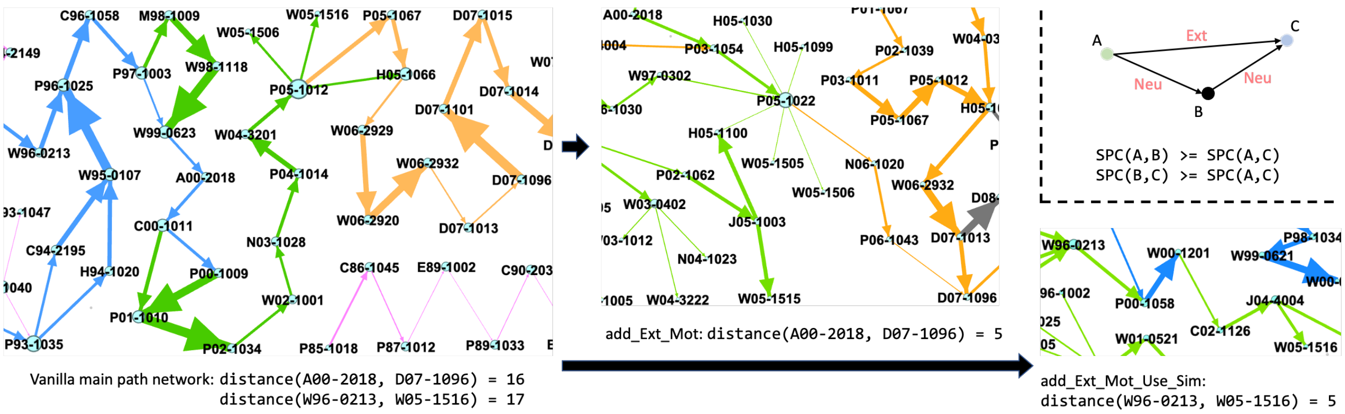

Figure 1. Motivations for Semantic Main Path Analysis.

To the best of our knowledge, this is the first paper which marries citation function classification to MPA. We proposed a systematic approach to semantic main path network extraction (Sect. 4) based on citation function classification (Sect. 3), which solves both issues raised above. Multiple semantic citation networks were built using different citation functions, for which multiple semantic main path networks were extracted, assuming that different semantic networks capture different types of knowledge flows between different knowledge entities, such as ideational basis, methodological extension, tool usage, and similarity in problem or methodology etc. We conjecture that different semantic main path networks will collectively provide a more comprehensive representation of an analysed domain. Note that, there were also some recent studies relying on citation importance classification (Hassan et al., 2018; Ghosal et al., 2022). Essentially, these approaches weighted citation edges by 1 (important) or 0 (incidental), screened out unimportant citations, did not further processing for knowledge flow analysis. The current paper is methodologically different. Citation function classification provides us with more flexible ways to perform MPA. The superiority of the proposed approach was qualitatively justified using two case studies (Sect. 5). To the best of our knowledge, this paper also proposed a three-way quantitative evaluation framework for the first time. Experiments proved that extracting and merging multiple semantic main path networks achieved better (topical) coverage, (topical) coherence and (ranking) pertinence (Sect. 6).

2. Related Work

2.1. Topological Approaches of Main Path Analysis

According to Verspagen (2007), MPA has two steps: citation weighting and main path extraction. Refer to Liu et al. (2019, 2020) for the discussions of best practices of each step. Citation weighting is traditionally based on each edge's traversal count in the search paths between a set of origin nodes and target nodes in a (usually reversed) citation network. We call them *topological* approaches. The ground-breaking work of Hummon and Dorerian (1989) defined three measures: Node Pair Project Count (NPPC), Search Path Link Count (SPLC), and Search Path Node Pair (SPNP). SLPC is predominantly used today. Bategelj (2003) proposed an efficient unified algorithm based on "standardising" citation networks (summarised in Table 1), and proposed the fourth measure Search Path Count (SPC). For each citation edge (u, v) in a standardized citation network, the citation weight is equal to the number of paths from pseudo-source to u multiplied by the number of paths from v to pseudo-sink. As citation networks are mostly acyclic, the calculation is done iteratively based on topological sort. Several adjustments exist. Liu and Kuan (2016) proposed to decay search path by length with the belief that knowledge diffusion has higher information loss along long paths, while Yu and Sheng (2021) used citing papers' citation influence for adjustment.

Table 1. Search Path Counting Methods for Main Path Analysis.

Method	Origins	Targets	Citation Network Standardization	
`NPPC`	All nodes	All	N/A	N/A
`SPLC`	All	Sinks (zero- outdegree)	Add a pseudo-source s^* and a pseudo-sink t^*	Connect s^* (resp. t^*) to all nodes (resp. sinks)
`SPNP`	All	All		Connect s^* and t^* to all nodes
`SPC`	Sources (zero-indegree)	Sinks		Connect s^* (resp. t^*) to all origins (resp. sinks)

Typically, main path extraction starts from certain chosen startpoints and greedily searches the highest weighted citation edges to follow. Verspagen (2007) enumerated paths from the source(s) with the maximal out-going edge weight as startpoint(s) so the main paths were called *forward local main paths* (Liu & Lu, 2012). Batagelj (2003) also tried the longest path as the *global main path* (Batagelj, 2003). Liu and Lu (2012) defined two new types of local main paths. *Backward local main path* starts from sinks and represents the significant knowledge flow from past to the most recent studies. They also found that these methods often miss the most significant citation edges, called *key-routes*, they proposed the fourth alternative called *key-route main path* which searches forward and backward simultaneously from key-routes. To increase the comprehensiveness of the extracted main paths, Liu and Lu (2012) heuristically selected the top-K startpoints or key-routes and merged the main paths extracted from them. Recently, Chen et al. (2022) proposed a more efficient dynamic programming algorithm for exhaustive main path extraction.

2.2. Semantic Approaches in Main Path Analysis

Liu et al. (2012) pioneered to use (expert-assigned) *citation relevancy* to adjust traversal count-based citation weighting. Of course, it be replaced by any semantic relatedness measure. For instance, Huang et al., (2022) claimed that using the weighted sum of the textual and structural similarities between cited and citing publications lead to better convergence, i.e., different slices of main path correspond well to different phases of domain development. Topic modelling is another popular semantic approach. Kim et al. (2022) used Latent Dirichlet Allocation (LDA) to analyse topic diffusion along main paths. Kim et al. (2018) used the Citation Influence Model, an extended LDA model which also models the generation process of each citing publication's citation mixture (Diez et al., 2007), to measure citation weights by topic similarity. Chen et al. (2022) calculated the Cosine similarity between citing and cited articles' topic distribution obtained by Latent Semantic Indexing (Deerwester et al., 1990). Notably, the citation relevancy of (u, v) is the sum of the pair-wise similarities between v and all other nodes u' on the current path towards v. While this treatment theoretically ensured the topical coherence of main path, it looks more straightforward to extract main paths from topic subnetworks and merge them. Community detection could be seen as an alternative way of finding topic subnetworks (Kim & Shin, 2018; Yu & Pan, 2021). To the best of our knowledge, citation function classification (Lyu et al., 2021; Kunnath et al., 2022) has never been applied to main path analysis before.

3. Citation Function Classification

3.1. Dataset and Annotation Schemes

We created a large citation function dataset by merging and re-annotating six existing datasets in the computational linguistics domain: `Teufel2010` (Teufel et al., 2006a; Teufel, 2010), `Dong2011` (Dong & Schäfer, 2011), `Jha2016` (Abu-Jbara et al., 2013; Jha et al., 2016), `Alvarez2017` (Hernández-Alvarez et al., 2017), `Jurgens2018` (Jurgens et al., 2018), and `Su2019` (Su et al., 2019). The source papers were crawled from ACL Anthology[2]. Different annotation guidelines were adopted so All citation contexts were-reannotated according to Teufel et al.'s 12-class annotation scheme (Teufel et al., 2006a) plus a "Future" class about future work. Reannotation is detailed in Supplementary Sect. A[3]. Some minority classes were still small, so we merged "PModi" with "PBas" into "Basis", and re-annotated "CoCo-" into "CoCoGM" or "CoCoRes". This resulted in our own 11-class annotation scheme, which was also mapped to 7-class and 6-class schemes by category merging. Table 2 shows the statistics of our dataset `Jiang2022`.

Table 2. Statistics of the Re-Annotated Dataset `Jiang2022` and Citation Function Scheme Mapping.

<table>
<tr><th colspan="4">Original Re-annotations (12+1 Class)</th><th colspan="3">Our 11-class Scheme**</th><th colspan="3">Mapped to 7-Class Scheme***</th><th colspan="3">Mapped to 6-class Scheme</th></tr>
<tr><th rowspan="2">label</th><th colspan="2">size</th><th>ratio</th><th rowspan="2">label</th><th>size</th><th rowspan="2">ratio</th><th rowspan="2">label</th><th>size</th><th rowspan="2">ratio</th><th rowspan="2">label</th><th>size</th><th rowspan="2">ratio</th></tr>
<tr><th>citstr</th><th>citseg*</th><th>citseg</th><th>citseg</th><th>citseg</th><th>citseg</th></tr>
<tr><td>Future</td><td>97</td><td>85</td><td>2.21%</td><td>Future</td><td>85</td><td>2.21%</td><td>Future</td><td>85</td><td>2.21%</td><td>Future</td><td>85</td><td>2.21%</td></tr>
<tr><td>CoCoXY</td><td>200</td><td>152</td><td>3.94%</td><td>CoCoXY</td><td>152</td><td>3.94%</td><td rowspan="3">Background</td><td rowspan="3">1773</td><td rowspan="3">46.00%</td><td rowspan="2">Background</td><td rowspan="2">1615</td><td rowspan="2">41.90%</td></tr>
<tr><td>Neut</td><td>1924</td><td>1463</td><td>37.96%</td><td>Neutral</td><td>1463</td><td>37.96%</td></tr>
<tr><td>Weak</td><td>223</td><td>158</td><td>4.10%</td><td>Weakness</td><td>158</td><td>4.10%</td><td rowspan="7">ComOrCon</td><td rowspan="7">944</td><td rowspan="7">24.49%</td></tr>
<tr><td>CoCoGM</td><td>390</td><td>299</td><td>7.76%</td><td rowspan="2">CoCoGM</td><td rowspan="2">328</td><td rowspan="2">8.51%</td><td rowspan="4">ComOrCon</td><td rowspan="4">479</td><td rowspan="4">12.43%</td></tr>
<tr><td rowspan="2">CoCo-</td><td rowspan="2">108</td><td rowspan="2">80</td><td rowspan="2">2.08%</td></tr>
<tr><td rowspan="2">CoCoRes</td><td rowspan="2">151</td><td rowspan="2">3.92%</td></tr>
<tr><td>CoCoR0</td><td>107</td><td>100</td><td>2.59%</td></tr>
<tr><td>PSup</td><td>123</td><td>100</td><td>2.59%</td><td>Support</td><td>100</td><td>2.59%</td><td rowspan="2">Similar</td><td rowspan="2">307</td><td rowspan="2">7.97%</td></tr>
<tr><td>PSim</td><td>247</td><td>207</td><td>5.37%</td><td>Similar</td><td>207</td><td>5.37%</td></tr>
<tr><td>PMot</td><td>365</td><td>288</td><td>7.47%</td><td>Motivation</td><td>288</td><td>7.47%</td><td>Motivation</td><td>288</td><td>7.47%</td><td>Motivation</td><td>288</td><td>7.47%</td></tr>
<tr><td>PUse</td><td>794</td><td>755</td><td>19.59%</td><td>Usage</td><td>755</td><td>19.59%</td><td>Uses</td><td>755</td><td>19.59%</td><td>Uses</td><td>755</td><td>19.59%</td></tr>
<tr><td>PModi</td><td>72</td><td>65</td><td>1.69%</td><td rowspan="2">Basis</td><td rowspan="2">167</td><td rowspan="2">4.33%</td><td rowspan="2">Extends</td><td rowspan="2">167</td><td rowspan="2">4.33%</td><td rowspan="2">Extends</td><td rowspan="2">167</td><td rowspan="2">4.33%</td></tr>
<tr><td>PBas</td><td>134</td><td>102</td><td>2.65%</td></tr>
<tr><td>Total</td><td>4784</td><td>3854</td><td></td><td></td><td>3854</td><td></td><td></td><td>3854</td><td></td><td></td><td>3854</td><td></td></tr>
</table>

* A citseg (citation segment) is a number of consecutive citstrs (citation string) cited in the same place. Citation function classification is done for each citseg.
** CoCoXY – Contrast/Comparison between two cited publications; CoCoGM/Res – Comparison/Contrast between cited and citing publications Goals or Methods/Results; Basis – Cited publication is ideationally based on; Support – Cited and citing publications support each other's claims or can be computationally plugged into each other.
*** ComOrCon – Comparison/Contrast between citing and cited publications.

[2] https://aclanthology.org/
[3] Supplementary materials: https://github.com/xiaoruijiang/CFC_MPN/blob/main/jasist2022_v2_SM_for_review.pdf

3.2. Citation Function Classification Models

For the purpose of recognizing citation functions more correctly, a series of deep learning models were developed. SciBERT was used to encode citation context, currently fixed to 2 and 3 sentences to each side of the citation sentence (*citance*). Three types of features were generated from the SciBERT-encoded context: (1) the *citation* representation **h**, from the citation segment (represented by a pseudo-word "CITSEG"), (2) the *citance representation*[4] **s**, pooled by citance encoder from the citation sentence, and (3) the *context representation* **c**, pooled by context encoder from the whole context. The final feature vector **f** was the concatenation of the three: **f** = [**h**; **s**; **c**]. Citation representation is mandatory because different citations in the same citance should have different feature representations, but citance and context representations were optional.

We tested two types of citation contexts. In a *sequential context*, no "[SEP]" (sequence separator) was inserted to separate context sentences. In this case, citance and context representations were directly pooled from citance tokens and context tokens respectively. Two options of citance encoder were tested: max-pooling and self-attention (Munkhdalai et al., 2016). In a *hierarchical context*, "[SEP]" symbols were inserted after each context sentence. Sentence representations were pooled using sentence pooler, for which "[SEP]" was used as the third option in addition to max-pooling and self-attention, and context representation was pooled indirectly from the representations of all context sentences. There were in total 34 model variants[5]. Due to the large GPU time required for training, we cherry-picked a subset of 11 relatively promising variants, shown in Table 3, based on initial experiments of all model variants with the 11-class scheme. Sect. 4.1 will discuss how to pick the appropriate models to perform semantic MPA based on per-class performance analysis of different models.

Table 3. Selected Citation Function Classification Models.

ID	citation_encoder (**h**)	context_type	sentence_pooler	citance encoder (**s**)	context_encoder (**c**)
1	O (used)	sequential	N/A	max_pool	max_pool
2-3	O	sequential	N/A	X (not used)	max_pool (2); self_attend (3)
4-6	O	sequential	N/A	max_pool (4); self_attend (5); X (6)	X
7-8	O	hierarchical	max_pool	X	max_pool (7); self_attend (8)
9-11	O	hierarchical	N/A	max_pool (9); self_attend (10); X (11)	X

[4] This choice was supported by the claim made by Lauscher et al. (2021) that most citation instances' functions could be determined only using citance alone.

[5] When **f** = **h**, depending on context_type, the number of model variants is 2. When **f** = [**h**; **s**], the number of model variants is: 2 (context_type = "sequential") + 2 × 3 (context_type = "hierarchical") = 8. When **f** = [**h**; **c**], if context_type = "sequential", the model variant number is 2; otherwise, if context_type = "hierarchical", it is 3 × 2 = 6 (3 sentence poolers by 2 context encoders). When **f** = [**h**; **s**; **c**], if context_type = "sequential", the model variant number is 2 × 2 (2 citance encoders multiplied by 2 context encoders) = 4; otherwise if context_type = "hierarchical", there are 2 × 3 × 2 = 12 model variants (2 citance encoders by 3 sentence poolers by 2 context encoders). Therefore, there are in total 2 + 8 + (2 + 6) + (4 + 12) = 34 model variants.

4. Semantic Main Path Analysis

4.1. Model Selection: Precision or Recall

Per-class performance analysis showed that no single best model could beat others on all citation functions or on all annotation schemes (Supplementary Table S1-S3). Therefore, we needed to choose the most appropriate model as a binary classifier for each specific citation function. The most pertinent citation function for MPA should be extension ("Basis"/"Extends") of cited work, and motivation ("Motivation") by previous studies. Figure 2- 3 show the performances of these two classes' top models. The darker the colour, the higher the performance. Although the best *extension* model was model 4 (seed = 5171, "seed =" omitted hereafter) with the 6-class scheme, its recall was less competitive. Considering the small size of the extension class, e.g., only 4.33% in our dataset, we decided to slightly **weigh recall over precision (recall-oriented)** and F1. The final choice had a good F1 and the highest recall, i.e., model 11 (47353, in solid red rectangle) trained with the 6-class scheme. Taking a similar recall-oriented approach, we chose model 7 (32491) trained with the 6-class scheme as the "best" *motivation* model.

We hoped that semantic citation networks could capture as many important citations as possible such as usage according to Valenzuella et al. (2015) and similarity according to Lu et al. (2014). For *usage* citations, we also took a recall-oriented approach. According to Figure 4, we opted for model 7 (13249) trained with the 11-class scheme which achieved the highest F1, because the recall of the chosen model was already high enough and its precision was much higher than other candidates. To further enrich the semantic citation network, we decided to add *similarity* citations because Teufel's annotation guidelines say similarity is between problems and solutions rather than results. According to Figure 5, the selected model was model 11 (25603) trained with the 11-class scheme.

The other way is to delete unimportant citations, e.g. neutral citations ("Neutral"/"Background") or future work citations ("Future") in our case. Due to the dominant size of neutral citations and high performance on this class (Figure 6), we decided to **trade recall for precision (precision-oriented)** for neutral ("Neutral"/"Background"), so model 2 (5171) with the 7-class scheme was selected. Because both precision and recall were high for future work citations (Figure 7), it was OK to adhere to the precision-oriented approach and select model 8 (32941) with the 11-class scheme because it achieved high enough precision and the best F1.

Extends		Jiang2021 (11-class)					Jiang2021 (7-class)					Jurgens2018 (6-class)				
ID	metric	5171	13249	25603	32491	47353	5171	13249	25603	32491	47353	5171	13249	25603	32491	47353
4	precision	70.37	65.52	63.33	62.50	74.07	65.52	76.00	63.64	70.97	64.71	80.77	56.76	52.94	62.50	68.97
	recall	55.88	55.88	55.88	58.82	58.82	55.88	55.88	41.18	64.71	64.71	61.76	61.76	52.94	58.82	58.82
	f1 score	62.30	60.32	59.38	60.61	65.57	60.32	64.41	50.00	67.69	64.71	70.00	59.15	52.94	60.61	63.49
8	precision	70.82	43.75	61.29	80.95	71.43	57.89	75.00	75.00	75.50	63.33	50.00	61.29	62.96	75.00	60.00
	recall	50.00	41.18	55.88	50.00	44.12	64.71	61.76	52.94	61.76	55.88	58.82	55.88	50.00	44.12	52.94
	f1 score	58.62	42.42	58.46	61.82	54.55	61.11	67.74	62.07	67.74	59.38	54.05	58.46	55.74	55.56	56.25
11	precision	53.85	61.29	70.83	55.88	63.33	72.00	64.71	61.09	57.58	57.89	65.52	50.00	72.00	53.85	67.65
	recall	61.76	55.88	50.00	55.88	55.88	52.94	64.71	55.88	55.88	64.71	55.88	41.18	52.94	61.76	67.65
	f1 score	57.53	58.46	58.62	55.88	59.38	61.02	64.71	58.46	56.72	61.11	60.32	45.16	61.02	57.53	67.65

Figure 2. Performances of Selected Models for Extension Citations.

Motivation		Jiang2022 (11-class)					Jiang2022 (7-class)					Jurgens2018 (6-class)				
ID	metric	5171	13249	25603	32491	47353	5171	13249	25603	32491	47353	5171	13249	25603	32491	47353
6	precision	57.14	60.27	63.79	56.72	60.66	59.32	67.21	62.90	71.15	56.06	62.96	58.33	60.94	55.07	53.52
	recall	68.97	75.86	63.79	65.52	63.79	60.34	70.69	67.24	63.79	63.79	56.82	60.34	67.24	65.52	65.52
	f1 score	62.50	67.18	63.79	60.80	62.18	59.83	68.91	65.00	67.27	59.68	60.71	59.32	63.93	59.84	58.91
7	precision	60.27	60.00	55.56	59.68	58.73	64.71	66.67	62.96	62.50	55.74	56.42	54.22	57.35	68.25	60.78
	recall	75.86	62.07	60.34	63.79	63.79	56.90	65.52	58.62	68.97	59.62	67.24	77.59	67.24	74.14	53.45
	f1 score	67.18	61.02	57.85	61.67	61.16	60.55	66.09	60.71	65.57	57.14	61.42	63.83	61.90	71.07	56.88
11	precision	62.90	60.00	66.67	68.85	54.41	66.67	67.27	58.57	57.14	58.90	54.41	66.13	69.84	61.29	74.55
	recall	67.24	62.07	68.97	72.41	63.79	68.97	63.79	70.69	68.97	74.14	63.79	70.69	75.86	65.52	70.69
	f1 score	65.00	61.02	67.80	70.59	58.73	67.80	65.49	64.06	62.50	65.65	58.73	68.33	72.73	63.33	72.57

Figure 3. Performances of Selected Models for Motivation Citations.

Usage		Jiang2022 (11-class)					Jiang2022 (7-class)					Jurgens2018 (6-class)				
ID	metric	5171	13249	25603	32491	47353	5171	13249	25603	32491	47353	5171	13249	25603	32491	47353
1	precision	73.65	73.33	74.48	75.84	71.52	74.50	75.71	78.29	76.80	73.20	74.65	79.07	76.06	80.95	78.69
	recall	72.19	80.13	71.52	74.83	74.83	73.51	70.20	66.89	63.58	74.17	70.20	67.55	71.52	56.29	63.58
	f1 score	72.91	76.58	72.97	75.33	73.14	74.00	72.85	72.14	69.57	73.68	72.35	72.86	73.72	66.41	70.33
2	precision	83.05	80.77	78.87	80.45	79.53	79.43	79.66	78.69	72.37	77.94	78.17	68.94	67.60	74.29	83.19
	recall	64.90	69.54	74.17	70.86	66.89	74.17	62.25	63.58	72.85	70.20	73.51	73.51	80.13	68.87	65.56
	f1 score	72.86	74.73	76.45	75.35	72.66	76.71	69.89	70.33	72.61	73.87	75.77	71.15	73.33	71.48	73.33
4	precision	79.39	79.67	77.08	74.64	78.17	78.99	75.54	70.55	73.10	79.85	82.03	81.54	72.79	83.74	76.67
	recall	68.87	64.90	73.51	68.21	73.51	72.19	69.54	76.16	70.20	70.86	69.54	70.20	70.86	68.21	76.16
	f1 score	73.76	71.53	75.25	71.28	75.77	75.43	72.41	73.25	71.62	75.09	75.27	75.44	71.81	75.18	76.41
7	precision	77.30	76.77	73.79	75.52	81.68	74.50	75.71	78.29	76.80	73.20	76.81	77.44	72.03	79.84	76.92
	recall	71.19	78.81	70.86	71.52	70.86	73.51	70.20	66.89	63.58	74.17	70.20	68.21	68.21	65.56	66.23
	f1 score	74.66	77.78	72.30	73.47	75.89	74.00	72.85	72.14	69.57	73.68	73.36	72.54	70.07	72.00	71.17

Figure 4. Performances of Selected Models for Usage Citations.

Similar		Jiang2022 (11-class)					Jurgens2018 (7-class)				
ID	metric	5171	13249	25603	32491	47353	5171	13249	25603	32491	47353
2	precision	57.45	46.94	65.79	60.00	67.50	62.07	53.12	54.29	59.65	63.46
	recall	64.29	54.76	59.52	71.43	64.29	58.06	54.84	61.29	54.84	53.23
	f1 score	60.67	50.55	62.50	65.22	65.85	60.00	53.97	57.58	57.14	57.89
5	precision	65.12	54.90	63.89	61.36	58.00	63.46	63.27	69.77	57.14	57.63
	recall	66.67	66.67	54.76	64.29	69.05	53.23	50.00	48.39	51.61	54.84
	f1 score	65.88	60.22	53.97	62.79	63.04	57.89	55.86	57.14	54.24	56.20
6	precision	58.14	56.82	58.54	62.79	60.53	59.32	70.00	60.71	66.67	81.08
	recall	59.52	59.52	57.14	64.29	54.76	56.45	45.16	54.84	58.06	48.39
	f1 score	58.82	58.14	57.83	63.53	57.50	57.85	54.90	57.63	62.07	60.61
11	precision	50.91	53.19	61.22	59.52	56.82	64.29	50.75	55.93	57.14	56.67
	recall	66.67	59.52	71.43	59.52	59.52	58.06	54.84	53.23	45.16	54.84
	f1 score	57.73	56.18	65.93	59.52	58.14	61.02	52.71	54.55	50.45	55.74

Figure 5. Performances of Selected Models for Similarity Citations.

Background		Jiang2022 (11-class)					Jiang2022 (7-class)					Jurgens2018 (6-class)				
ID	metric	5171	13249	25603	32491	47353	5171	13249	25603	32491	47353	5171	13249	25603	32491	47353
2	precision	75.50	76.14	76.04	78.85	75.43	82.17	81.27	78.02	80.72	78.59	78.06	76.74	75.08	74.34	74.34
	recall	77.82	68.60	74.74	69.97	75.43	82.87	79.21	81.74	82.30	78.37	76.85	78.40	74.38	77.78	77.78
	f1 score	76.64	72.17	75.39	74.14	75.43	82.52	80.23	79.84	81.50	78.48	77.45	77.56	74.73	76.02	76.02
6	precision	76.36	71.78	72.78	78.75	75.77	82.07	82.91	80.11	79.94	76.07	78.22	76.38	76.88	73.08	76.56
	recall	71.67	79.86	78.50	77.13	75.77	75.84	81.74	83.71	80.62	84.83	78.70	76.85	75.93	82.10	75.62
	f1 score	73.94	75.61	75.53	77.93	75.77	78.83	82.32	81.87	80.28	80.21	78.46	76.62	76.40	77.33	76.09
8	precision	75.60	73.61	75.95	78.07	69.64	79.53	76.49	78.17	80.50	76.62	76.05	76.98	72.30	76.88	71.43
	recall	75.09	72.35	75.43	71.67	79.86	86.24	86.80	81.46	81.18	82.87	72.53	69.14	76.54	75.93	78.70
	f1 score	75.34	72.98	75.68	74.73	74.40	82.75	81.32	79.78	80.84	79.62	74.25	72.85	74.36	76.40	74.89
11	precision	74.38	73.11	76.22	73.75	75.68	77.15	78.98	82.20	77.40	81.65	67.94	76.20	75.00	76.19	75.37
	recall	82.35	76.11	79.86	75.77	76.45	88.20	78.09	77.81	83.71	75.00	82.41	78.09	80.56	74.07	79.32
	f1 score	78.12	74.58	78.00	74.75	76.06	82.31	78.53	79.94	80.43	78.18	74.48	77.13	77.68	75.12	77.29

Figure 6. Performances of Selected Models for Neutral/Background Citations.

Future		Jiang2022 (11-class)					Jiang2022 (7-class)					Jurgens2018 (6-class)				
ID	metric	5171	13249	25603	32491	47353	5171	13249	25603	32491	47353	5171	13249	25603	32491	47353
2	precision	87.50	84.62	66.67	76.47	72.22	75.00	70.59	92.31	72.22	76.47	75.00	100.00	87.50	81.25	76.47
	recall	82.35	64.71	94.12	76.47	76.47	70.59	70.59	70.59	76.47	76.47	70.59	64.71	82.35	76.47	76.47
	f1 score	84.85	73.33	78.05	76.47	74.29	72.73	70.59	80.00	74.29	76.47	72.73	78.57	84.85	78.79	76.47
5	precision	82.35	88.24	68.42	73.68	81.25	86.67	83.33	72.22	92.86	86.67	92.31	72.22	81.25	85.71	100.00
	recall	82.35	88.24	76.47	82.35	76.47	76.47	58.82	76.47	76.47	76.47	70.59	76.47	76.47	70.59	82.35
	f1 score	82.35	88.24	72.22	77.78	78.79	81.25	68.97	74.29	83.87	81.25	80.00	74.29	78.79	77.42	90.32
7	precision	76.47	65.00	92.86	72.22	85.71	76.47	80.00	68.75	77.78	73.33	68.42	68.42	68.42	86.67	100.00
	recall	76.47	76.47	76.47	76.47	70.59	76.47	70.59	64.71	82.35	64.71	76.47	76.47	76.47	76.47	76.47
	f1 score	76.47	70.27	83.87	74.29	77.42	76.47	75.00	66.67	80.00	68.75	72.22	72.22	72.22	81.25	86.67
8	precision	92.86	81.25	80.00	93.75	92.86	80.00	66.67	76.47	66.67	82.35	86.67	81.25	76.47	65.00	81.25
	recall	76.47	76.47	70.59	88.24	76.47	70.59	82.35	76.47	82.35	82.35	76.47	76.47	76.47	76.47	76.47
	f1 score	83.87	78.79	75.00	90.91	83.87	75.00	73.68	76.47	73.68	82.35	81.25	78.79	76.47	70.27	78.79

Figure 7. Performances of Best Models for Future Work Citations.

4.2. Semantic Main Path Network Extraction

4.2.1. Citation Network Building

Starting from an empty citation network, a citation edge was added between a pair of publications if there existed **at least one** in-text citation about extension **or** motivation (`add_Ext_Mot`) using the "best" *extension* or *motivation* models selected in the recall-oriented approach in Sect. 3.1. Taking the same recall-oriented approach, more citation edges were added if there existed at least one *usage* citation (`plus_add_Use`), and the semantic citation network was further expanded with *similarity* citations (`plus_add_Sim`). On the other hand, we also built the fourth semantic citation network by deleting unimportant in-text citations from the original citation network. For each pair of publications, if **all** in-text citations between them were *neutral* or *future work* citations, the citation edge was removed from the citation network (`del_Bkg_Fut`).

4.2.2. Main Path Network Extraction

The semantic citation networks we analysed have many small strongly connected components (SCC), so we applied the Simple Search Path Count approach (Jiang et al., 2020), an extension of SPC to deal with cyclic citation networks, for MPN extraction. Their `JMPA` package[6] (Java package for MPA) was used for implementation. Following Jiang et al. (2020), we segmented the network under analysis to several time slices, extracted top-K (K = 10) key-route main paths (Liu et al., 2012) from each slice, and merged them into an MPN. Supplementary Sect. B details the experimental setup.

---

[6] https://github.com/xiaoruijiang/JMPA

5. Qualitative Analysis

For experimental analysis, citation data came from the 2015 version of ACL Anthology Network (AAN; Radev et al., 2013) about computational linguistics/natural language. Three areas were selected: natural language parsing[7] (`AANPar`), automatic document summarisation (`AANSum`), and machine translation (`AANMT`). Due to space limit, this section showcases on `AANPar` and `AANSum` to demonstrate the superiority of semantic MPA. Supplementary Table S4 summarises the statistics of the (semantic) citation networks and their time slices. The experimental setup is detailed in Supplementary Sect. B. Key-route MPA was used for main path extraction. Because other connected components are all small islands in the citation network, it was valid to follow the common practice in MPA to extract semantic MPNs from the largest connected component (CC).

5.1. Case Study 1: Natural Language Parsing

5.1.1. Main Path Network

For comparison purpose, Figure 8 presents the MPN extracted from the original citation network `AANPar`. Topic branches are numbered. Seminal papers (verified according to the authors' knowledge about the domain) are in red rectangles, while survey-style papers are in ovals, such as special issue or shared task introduction papers. Table 4 shows a subset of representative main path papers on each topic branch and Supplementary Table S5 presents the complete list. Topic keywords and short excerpts for certain papers are to assist understanding. Branch 1 describes the early studies about various grammatical formalisms[8], such as **categorical grammar**, **unification grammar**, **categorical unification grammar**, and **Lambek calculus** etc. However, since late 1980s, the domain started to have a sense of probabilistic thinking (Branch 2). Branch 3 shows the important development where **Penn TreeBank** (J93-2004 and H94-1020) appeared as the most important linguistic resource that most future papers used for developing and evaluating parsing techniques.

Branch 4 represents the mainstream of statistical parsing in the 1990s and 2000s, such as **maximum entropy modelling** (W96-0213, A00-2018) or in another name **log-linear model** (P04-1014), **conditional random fields** (N03-1028), and **max-margin parsing** (W04-3201, P05-1012). Note that, C00-1011 and P00-1009 were two papers on **Data-Oriented Parsing (DOP)** promoted by Rens Bod, which however ceased in the wave of statistical parsing dominated by other proposals presented above. Early studies about **dependency analysis** blossomed into the huge Branch 5 and became the dominant trend since

---

[7] Parsing: Parsing, syntax analysis, or syntactic analysis is the process of analyzing a string of symbols, either in natural language, computer languages or data structures, conforming to the rules of a formal grammar. See Wikipedia page: https://en.wikipedia.org/wiki/Parsing.
[8] Note that more grammars were proposed even earlier, outside our time range of analysis.

around 2005, further expediated by two important shared tasks W06-2920 and D07-1096, which then diverted into Branch 6 about dependency parsing of **morphologically rich languages** and Branch 7 about **cross-lingual dependency parsing**. An issue was that many main path papers were connected by incidental citations. For instance, the citation from A00-2018 said that C00-1011 "stays behind the scores of" the former, a weak citation about performance comparison. For another instance, H91-1037 received only 10 citations in our dataset. `SPC`(H91-1037, J93-2004) was high only because of high-impact citing citing paper J93-2004 (1006 citations), although the citation was incidental.

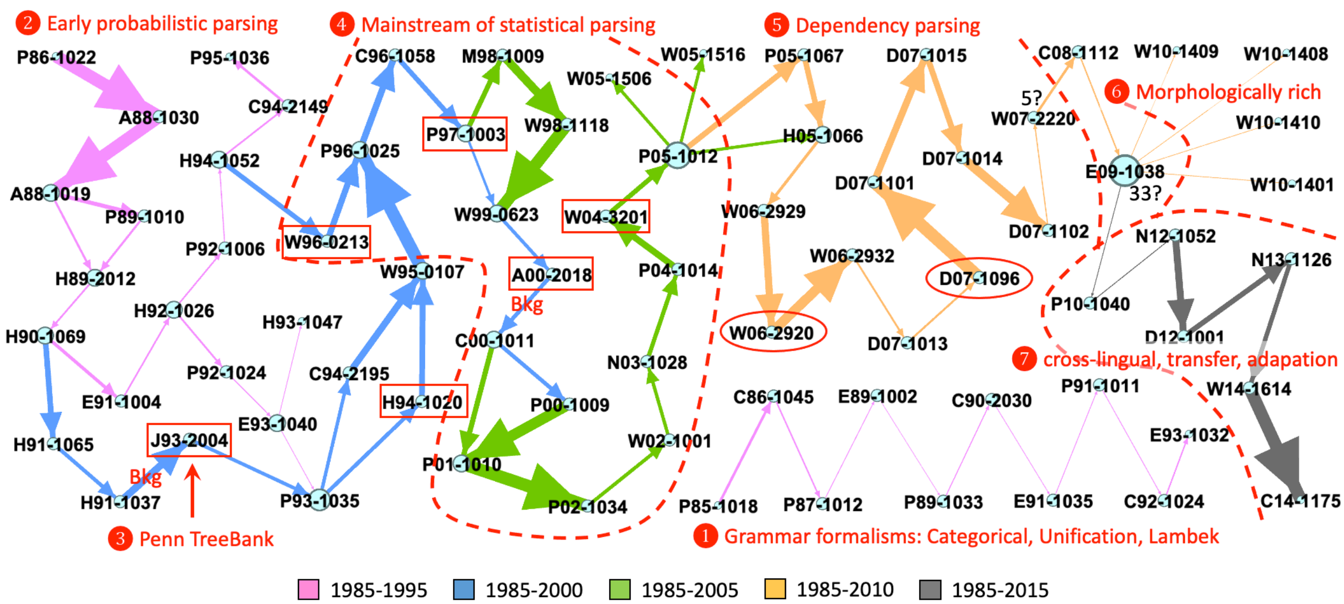

Figure 8. Main Path Network Extracted from `AANPar`.

Table 4. Representative Main Path Papers Extracted from `AANPar`.

ACLID	Title
Branch 1	
C86-1045	**Categorial Unification Grammars**
P87-1012	A Lazy Way To Chart-Parse With **Categorial Grammars**
C90-2030	Normal Form Theorem Proving For The **Lambek Calculus**
Branch 2	
H92-1026	Towards History-Based Grammars: Using Richer Models For **Probabilistic Parsing**
E93-1040	Parsing The Wall Street Journal With The Inside-Outside Algorithm Excerpt: We report grammar inference experiments on partially parsed sentences taken from the **Wall Street Journal corpus** using the inside-outside algorithm for **stochastic context-free grammars**.
Branch 3	
J93-2004	**Building A Large Annotated Corpus Of English: The Penn Treebank**
H94-1020	**The Penn Treebank: Annotating Predicate Argument Structure**
Branch 4	
A00-2018	A **Maximum-Entropy-Inspired** Parser
C00-1011	Parsing With The Shortest Derivation (about **DOP by** Rens Bod)
P00-1009	An Improved Parser For **Data-Oriented** Lexical-Functional Analysis (about **DOP by** Rens Bod)
N03-1028	Shallow Parsing With **Conditional Random Fields**
P04-1014	Parsing The WSJ Using **CCG** And **Log-Linear Models**
W04-3201	**Max-Margin Parsing**
Branch 5	
W06-2920	**CoNLL-X Shared Task On Multilingual Dependency Parsing**
D07-1096	**The CoNLL 2007 Shared Task on Dependency Parsing**
D07-1014	Probabilistic Models of **Nonprojective Dependency Trees**
Branch 6	
W10-1401	Statistical Parsing of **Morphologically Rich** Languages (SPMRL) What How and Whither
W10-1410	Lemmatization and Lexicalized Statistical Parsing of **Morphologically-Rich** Languages: the Case of **French**
Branch 7	
N12-1052	**Cross-lingual** Word Clusters for Direct **Transfer** of Linguistic Structure
N13-1126	**Target Language Adaptation** of Discriminative **Transfer** Parsers

5.1.2. Semantic Main Path Network: Add Extension and Motivation Citations

The above observations motived us to exploit the semantic relationships between papers in MPA. Figure 9-12 show the semantic MPNs extracted from the four semantic citation networks induced from `AANPar`, namely `AANPar_add_Ext_Mot`, `AANPar_plus_add_Use`, `AANPar_plus_add_Sim`, and `AANPar_del_Bkg_Fut`. Interesting chemical reactions occurred when MPA met citation function classification. Each semantic MPN revealed some novel branches or new papers. They collectively drew a more comprehensive picture of domain development. Supplementary Sect. D presents selected citation context excerpts to help readers understand the citation functions marked on certain edges.

On `AANPar_add_Ext_Mot` (Figure 9 and Table 5, and Supplementary Table S6 for a complete list of main path papers), the early development of parsing technology was tected. Branch 2 is a new branch about old parsers such as **shift-reduce parsing**, **left-corner parsing**, **tabular parsing**, and **left-to-right (LR) parsing** etc. Similarly, we saw another (isolated) early development of **probabilistic** approaches (Branch 3; details in Supplementary Table S6). In addition to A00-2018 as the source of the statistical parsing mainstream, a third source started from E85-1024 ("A probabilistic parser") to J94-2001 ("Tagging English Text with a Probabilistic Model") and W96-0213, then through P02-1034 into the new Branch 4 about **multiple parse ranking and re-ranking**. Note that Branch 5 started went into a "dead" end about "Chinese TreeBank" (W00-1201).

From the right part of Figure 9, we saw a branch of **DOP** papers published by Rens Bod until P01-1010. Similar to the evolution pathway in Figure 8, it was gradually merged into the dominant dependency parsing branch. D08-1059 ("A Tale of Two Parsers: Investigating and Combining Graph-based and Transition-based Dependency Parsing") was motivated (denoted by "Mot" on the edge) by two papers P07-1050 ("K-best Spanning Tree Parsing") and D07-1013 ("Characterizing the Errors of Data-Driven Dependency Parsing Models").

Note that, there was a potentially problematic Branch 8 about machine translation (MT) using dependency parsing. Concerning (P05-1012, H05-1066), the citation context excerpt below reveals that although "improving upon" may indicate an extension, the whole context may be recognised as "Similar" or "CoCoGM". This shows that multi-label classification might be a promising future direction to explore (Lauscher at al., 2021).

> *"We mentioned above that our approach appears to be similar to that of reranking for statistical parsing (Collins, 2000; Charniak and Johnson, 2005). While it is true that we are improving upon the output of the automatic parser, we are not considering multiple alternate parses."*

Vague cases exist, such as (W00-1201, C02-1126), a self-citation by D. M. Bikel and D. Chiang. From the citation context excerpt below, expressions like "starting from" and "we have modified" might have been selected as strong signals for extension class ("Ext").

"The third experiment was on the Chinese Treebank, starting with the same head rules used in (Bikel and Chiang, 2000). These rules were originally …, and although we have modified them for parsing, …"

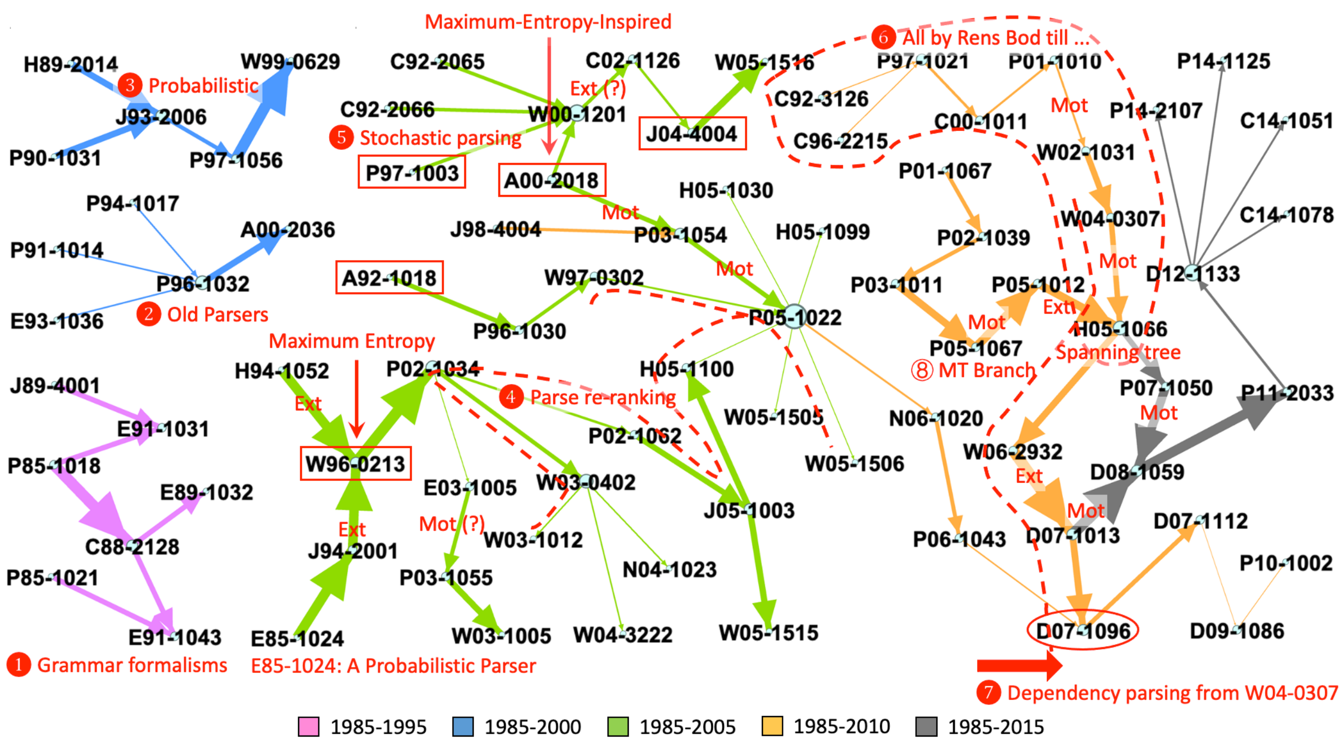

Figure 9. Main Path Network Extracted from `AANPar_Ext_Mot`.

Table 5. Representative Main Path Papers Extracted from `AANPar_add_Ext_Mot`.

ACLID	Title
Branch 2	
P91-1014	Polynomial Time And Space **Shift-Reduce Parsing** Of Arbitrary **Context-Free Grammars**
E93-1036	Generalized **Left-Corner Parsing**
P94-1017	An Optimal **Tabular Parsing** Algorithm
Branch 4	
W97-0302	Global Thresholding And **Multiple-Pass Parsing**
P02-1034	New **Ranking Algorithms** For Parsing And Tagging: Kernels Over Discrete Structures And The Voted Perceptron
P05-1022	Coarse-To-Fine **N-Best Parsing** And MaxEnt **Discriminative Reranking**
…	
Branch 5	
C92-2065	**Probabilistic Tree-Adjoining Grammar** As A Framework For **Statistical** Natural Language Processing
C92-2066	**Stochastic Lexicalized Tree-Adjoining Grammars**
…	
W00-1201	Two Statistical Parsing Models Applied To The ***Chinese Treebank***
Branch 6	(a dead branch)
C92-3126	A Computational Model Of Language Performance: ***Data Oriented Parsing***
P97-1021	A ***DOP*** Model For Semantic Interpretation
Branch 8	(A “wrong” branch)
P01-1067	A *Syntax-Based Statistical Translation Model*
P03-1011	Loosely *Tree-Based* Alignment For *Machine Translation*
P05-1067	*Machine Translation* Using Probabilistic Synchronous **Dependency** Insertion Grammars

5.1.3. Semantic Main Path Network: Further Add Usage and Similarity Citations

By further adding usage citations, i.e., on `AANPar_plus_add_Use`, we saw drastically richer diversity in the development branches (Figure 10, Table 6 and Supplementary Table S7). Again, statistical parsing techniques evolved from multiple intelligent sources (Branches 1-3). A clear notion of "**corpus-based**" parsing emerged (Branch 1). Branch 2 was motivated by H93-1047 ("Automatic Grammar Induction And Parsing Free Text: A Transformation-Based Approach", a duplicate of P93-1035) and developed into "**shallow parsing**" of words into "**text chunks**"[9]. This time, the seminal paper J93-2004 about the **Penn Treebank** project emerged in Branch 3 and developed through W96-0213 to J04-4004. Most subsequent papers used Peen Treebank for development and evaluation. We also saw a similar evolution into the start of the wrong MT branch P01-1067, a paper on syntax-based statistical translation. The citation context excerpt below shows that P01-1067 *used* J95-4004.

> *"Brill's part-of-speech (POS) tagger (Brill, 1995) and Collins' parser (Collins, 1999) were used to obtain parse trees for the English side of the corpus."*

The DOP branch lead by Rens Bod from J93-2006 into C00-1011 (Branch 6 in Table 5) "developed" into Branch 4 and found the important shared task W05-0620 on **semantic role labelling** (SRL) of predicate arguments, and "vanished". This is understandable because SRL became a rather standalone area since then[10] and began to cite less and be less cited by parsing papers. In addition, the branch about **cross-lingual dependency parsing** embraced a more diverse set of papers.

By further adding similarity citations, i.e., on `AANPar_plus_add_Sim`, the semantic MPN bore high similarity (Figure 11, Table 7 and Supplementary Table S8). However, we observed that quite a few interesting new branches emerged. Starting from the seminal Penn Treebank paper J93-2004, two new branches developed from P97-1062 and W97-0301 based on *usage* citations respectively into Branch 1 about **rhetorical parsing** and Branch 2 about **probabilistic parsing with CCG (Combinatory Categorial Grammar)**. Through *similarity* citations, we found some new main path papers, such as J96-1002 ("A Maximum Entropy Approach to Natural Language Processing") which was heavily cited (387 times). The following citation context excerpt proved that similarity citation is indeed relevant to knowledge flow of scientific ideas.

> *"The maximum entropy models used here are similar in form to those in (Ratnaparkhi, 1996; Berger, Della Pietra, and Della Pietra, 1996; Lau, Rosenfeld, and Roukos, 1993)."*

---

[9] From Wikipedia, shallow parsing is also chunking or light parsing: https://en.wikipedia.org/wiki/Shallow_parsing

[10] Both semantic role labelling and dependency parsing became rather standalone topics and had bespoke monographs on these two topics.

The domain then evolved to the dominant dependency parsing branch (Branch 3), where we were excited to see two new shared tasks about **joint syntactic and semantic dependency parsing** (W08-2121, W09-1201), and then to Branch 4 of subsequent studies on **semantic dependency parsing** (W09-1208, D09-1004).

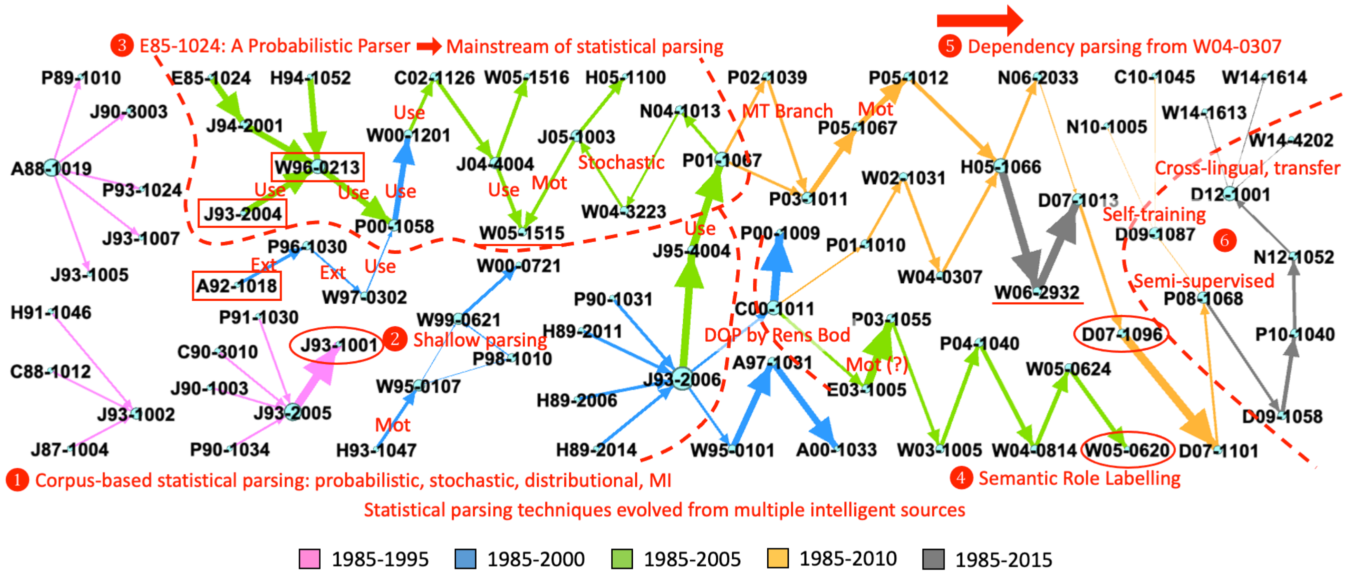

Figure 10. Main Path Network Extracted from `AANPar_plus_add_Use`.

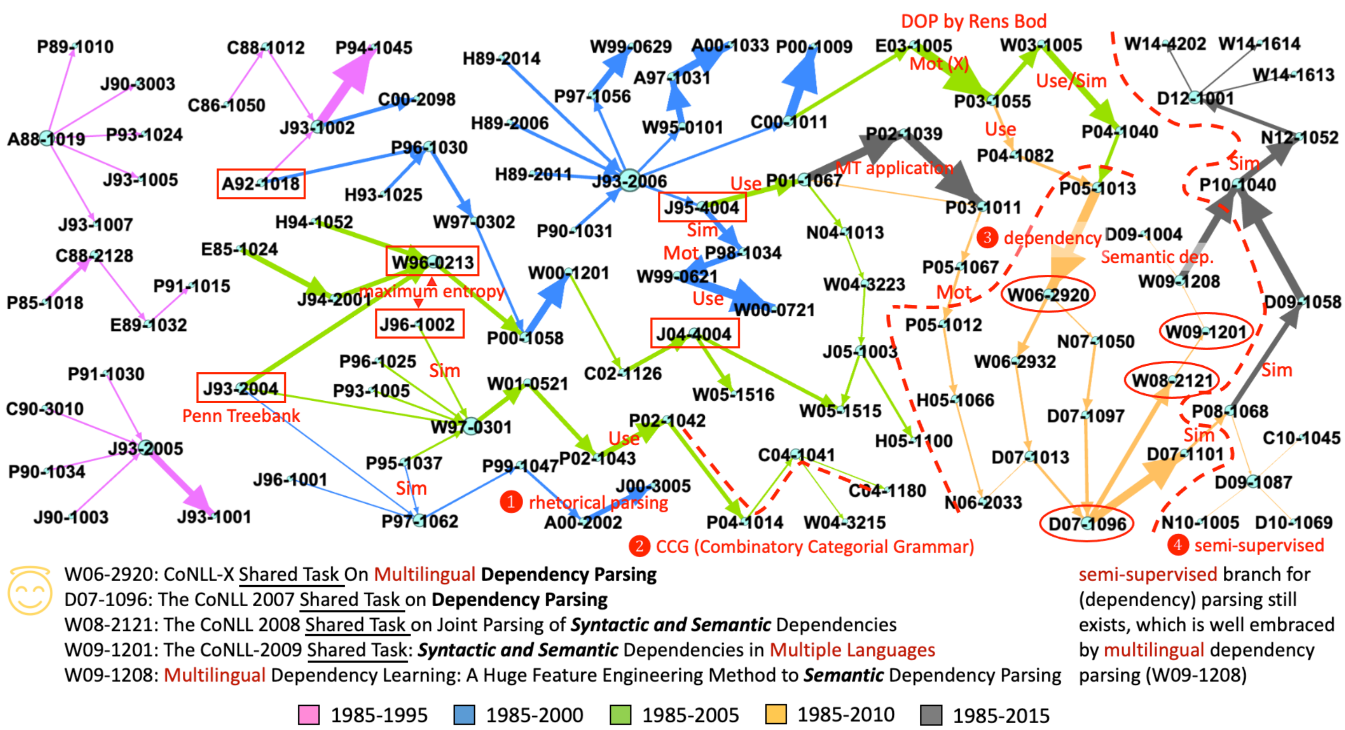

Figure 11. Main Path Network Extracted from `AANPar_plus_add_Sim`.

Table 6. Representative Main Path Papers Extracted from `AANPar_plus_add_Use`.

ACLID	**Title**
Branch 1	
J93-1002	Generalized **Probabilistic LR Parsing** Of Natural Language (**Corpora**) With Unification-Based Grammars
J93-1001	**Introduction To The Special Issue On Computational Linguistics Using Large Corpora**
P90-1031	Parsing The LOB **Corpus**
Branch 2	
W95-0107	**Text Chunking** using Transformation-Based Learning
W00-0721	**Shallow Parsing** by Inferencing with Classifiers
Branch 3	
…	
J93-2004	**Building A Large Annotated Corpus Of English: The Penn Treebank**
…	
W00-1201	Two **Statistical Parsing** Models Applied To The ***Chinese Treebank***
…	
P01-1067	A Syntax-Based *Statistical Translation Model*
Branch 4	
W05-0620	**Introduction To The CoNLL-2005 Shared Task: Semantic Role Labeling**
Branch 6	(Extended branch about cross-lingual dependency parsing)
P08-1068	Simple **Semi-supervised** Dependency Parsing
D09-1087	**Self-Training** PCFG Grammars with Latent Annotations Across Languages
W14-1613	Distributed Word Representation Learning for **Cross-Lingual** Dependency Parsing

Table 7. Representative Main Path Papers Extracted from `AANPar_plus_add_Sim`.

ACLID	**Title**
Branch 1	
P99-1047	A Decision-Based Approach To **Rhetorical Parsing**
J00-3005	The **Rhetorical Parsing** Of Unrestricted Texts: A Surface-Based Approach
Branch 2	
P02-1042	Generative Models For **Statistical Parsing** With **Combinatory Categorial Grammar**
P04-1014	Parsing The WSJ Using **CCG** And **Log-Linear Models**
C04-1180	Wide-Coverage Semantic Representations From A **CCG Parser**
Branch 3	(extended branch of dependency parsing)
…	
W08-2121	**The CoNLL 2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies**
W09-1201	**The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages**
Branch 4	(extended to semantic dependency parsing)
W09-1208	**Multilingual Dependency Learning**: A Huge Feature Engineering Method to **Semantic Dependency Parsing**
D09-1004	**Semantic Dependency Parsing** of **NomBank** and **PropBank**: An Efficient Integrated Approach via a Large-scale Feature Selection

5.1.4. Semantic Main Path Network: Delete Neutral and Future Work Citations

Finally, on `AANPar_del_Bkg_Fut` (Figure 12, Table 8 and Supplementary Table S9), we observed some interesting branches or papers. Since P08-1068, the domain diverted into a new branch about **optimization techniques** used in parsing algorithms, such as **dynamic programming**, **integer linear programming** and **dual decomposition** (Branch 2). Branch 3 was a similar cross-lingual dependency parsing branch, but it evolved into Branch 4 about parsing morphologically rich languages through a new **shared task** (W13-4917), thus provided a complementary view to Branch 6 in Figure 8. We postulate the result is meaningful since dependency parsing was directed by important shared tasks. Note that, deleting neutral and future work citations might result in weaker semantic coherence than by adding more significant citations like extension and similarity (quantified in Sect. 6.3). For example, N07-1069 only made a result comparison with W06-2928, therefore it is less confident to say scientific ideas flew through this path.

> *"Here we can compare directly with the best systems for this dataset in CoNLL-X. The best system (Corston-Oliver & Aue, 2006), ..."*

In summary, we conjecture that multiple semantic MPNs extracted from different types of semantic citation networks reveal complimentary views and novel knowledge flows, thus should be merged into a more comprehensive representation of scientific domain's topic evolution.

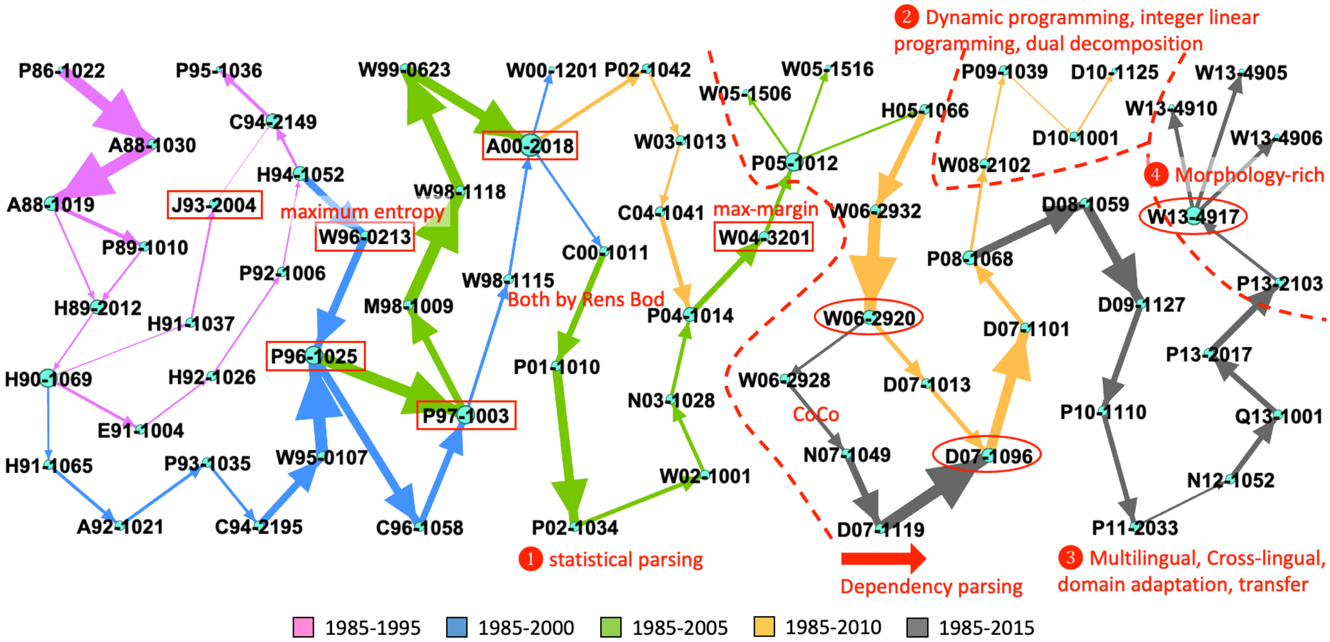

Figure 12. Main Path Network Extracted from `AANPar_del_Bkg_Fut`.

Table 8. Representative Main Path Papers Extracted from `AANPar_del_Bkg_Fut`.

ACLID	Title
Branch 2	
W08-2102	TAG, **Dynamic Programming**, and the Perceptron for Efficient, Feature-Rich Parsing
P09-1039	Concise **Integer Linear Programming** Formulations for Dependency Parsing
D10-1001	On **Dual Decomposition** and **Linear Programming** Relaxations for Natural Language Processing
Branch 3	
W06-2928	Dependency Parsing With Reference To Slovene **Spanish** And **Swedish**
D07-1119	**Multilingual** Dependency Parsing and **Domain Adaptation** using DeSR
P13-2017	Universal Dependency Annotation for **Multilingual** Parsing
Branch 4	
W13-4917	**Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages** W13-4905, W13-4906 and W13-4910 are all SPMRL 2013 Shared Task papers

5.2. Case Study 2: Automatic Document Summarisation

Due to space limit, an informative summary is presented here (Figure 13-17). See Tables S10-S14 in Supplementary Sect. E for the details of main path papers and Supplementary Sect. F for citation context excerpts. The MPN extracted from `AANSum` (Figure 12) covered a few early summarization studies centering around the usage of **semantic coherence** devices (Branch 1), such as **discourse structure**, **rhetorical relations**, and **lexical chains** (W97-0703: Using Lexical Chains For Text Summarization) etc. Then the main body of literature focused on **multidocument summarization** (Branch 2) pioneered by the seminal journal article J98-3005 ("Generating Natural Language Summaries From Multiple On-Line Sources"). The subsequent studies in this topic eventually gave birth to an important **Special Issue** on Summarization (J02-4001). Since the advent of PageRank in 1998, the **graph-based ranking** idea was introduced to the summarization domain for sentence ranking for extractive summarisation (Branch 3). Seminal works included P04-3020 ("Graph-Based Ranking Algorithms For Sentence Extraction Applied To Text Summarization"), W04-3252 ("TextRank: Bringing Order Into Texts"), the subsequent demonstration paper of TextRank (P05-3013), and its extension to multidocument summarization (I05-2004). More recently, a large body of the literature were about some interleaved topics: **optimization techniques** such as **submodular optimization** (E12-1023), **integer linear programming** (D12-1022), and **dual decomposition** (P13-1020); **compressive summarization** (P10-1058, P11-1049, D13-1047); and **compressive summarization based on dependency tree** (P14-2052, D14-196). Notably, comparison (sometimes weakness) function was the dominating citation function in Branch 4 in Figure 12. In addition, the only papers about summarization evaluation was N03-1020 about ROUGE ("Automatic Evaluation Of Summaries Using N-Gram Co-Occurrence Statistics"). These two drawbacks motivated us to explore novel branches of summarization using semantic MPNs.

By adding extension and motivation citations (Figure 14), we could see a larger early branch about the usage of rhetorical structure and found a seminal application in **scientific summarization** (J02-4002), which was extended by subsequent studies in other areas, like W03-0505 ("Summarising Legal Texts: Sentential Tense And Argumentative Roles"), evidenced by the citation context excerpt below.

> *"Our methodology builds and extends the Teufel and Moens (Teufel and Moens, 2002) approach to automatic summarization."*

In addition to the common topics like multidocument summarization (Branch 2) and graph-based ranking algorithms (Branch 5), we were also excited to see Branch 3 about **automatic evaluation** and related studies. Heavily cited ones included

N03-1020 and W04-1013 about the ROUGE package. We also saw more studies about **sentence reduction**, **compression and fusion** for summarization. Both Branch 4-1 and 4-2 were pioneered by K. R. McKewon in A00-1043 ("Sentence Reduction For Automatic Text Summarization"), A00-2024 ("Cut and Paste Based Text Summarization"), and J05-3002 ("Sentence Fusion For Multidocument News Summarization").

By further adding usage citations (Figure 15), although we lost the graph-based ranking branch (despite that we got a new paper W04-3247 about LexPageRank), we could uncover more novel topics and branches. Branch 2 about automatic evaluation included more important papers such as N04-1019 about the **Pyramid method** ("Evaluating Content Selection In Summarization: The Pyramid Method"). A significant new branch was Branch 3 about **scientific summarization** at right bottom, starting from the seminal paper J02-4002 to **citation function classification** (W06-1613, N07-1040) and **citation-based summarization** (C08-1087, N09-1066, P10-1057, and C10-1101). By further adding similarity citations (Figure 16), we could see one obvious expansion of Branch 1 about evaluation, starting from **factoid analysis** (W04-3254) to summarization **evaluation without human models**, including D09-1032 ("Automatically Evaluating Content Selection in Summarization without Human Models") and C10-2022 ("Multilingual Summarization Evaluation without Human Models"), both written by famous researchers in this domain (A. Nenkova and H. Saggion respectively).

Finally, the MPN extracted from `AANSum_del_Bkg_Fut` (Figure 17) recovered the vanished or shrunk branches about **multidocument summarization** (Branch 1) and **graph-based ranking** (Branch 2), and at the same time introduced some new papers, such as C04-1129 for Branch 1 ("Syntactic Simplification For Improving Content Selection In Multi-Document Summarization"), P08-1048 for Branch 2 ("Summarizing Emails with Conversational Cohesion and Subjectivity", whose abstract says "Second, we use two graph-based summarization approaches, …, to extract sentences as summaries."), and W09-1802 ("A Scalable Global Model for Summarization", whose abstract says "We present an Integer Linear Program for … for automatic summarization.") and C10-2105 ("Opinion Summarization with Integer Linear Programming Formulation for Sentence Extraction and Ordering") for Branch 3 about optimisation methods for summarisation.

Again, by gradually adding more citation semantics, the semantic MPNs together proved to be more expressive than the semantics-agnostic counterpart.

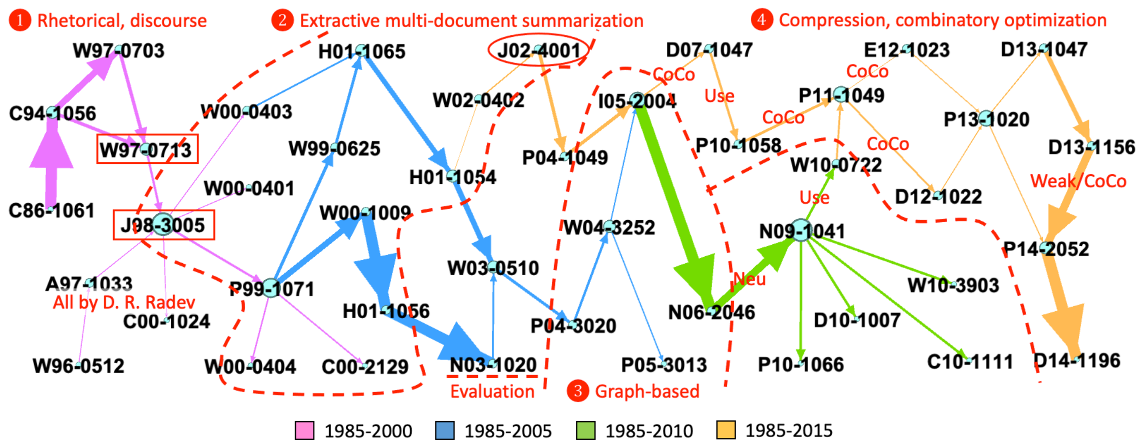

Figure 13. Main Path Network of the Summarisation Network `AANSum`.

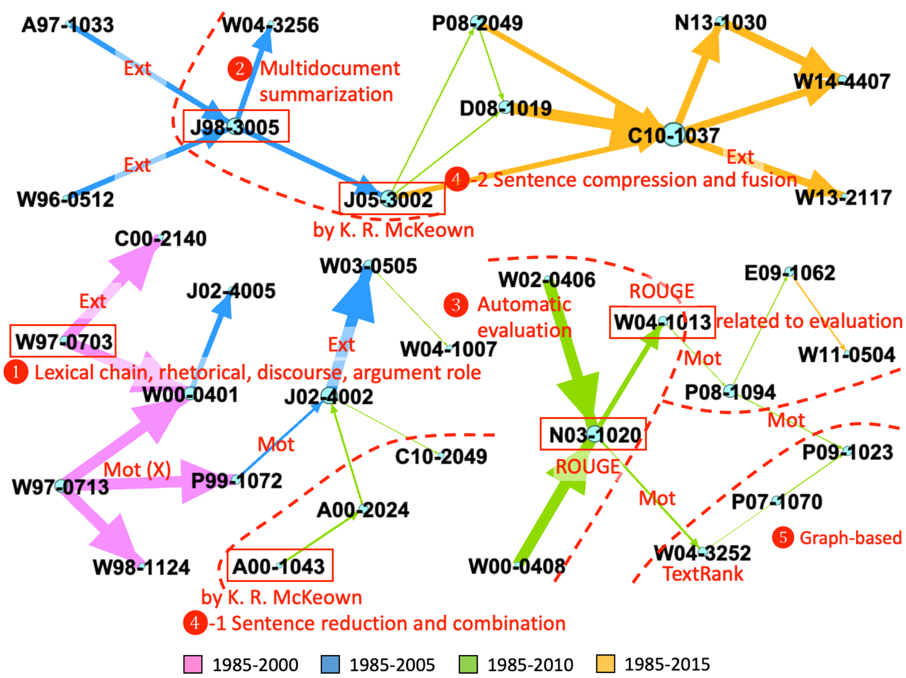

Figure 14. Main Path Network Extracted from `AANSum_add_Ext_Mot`.

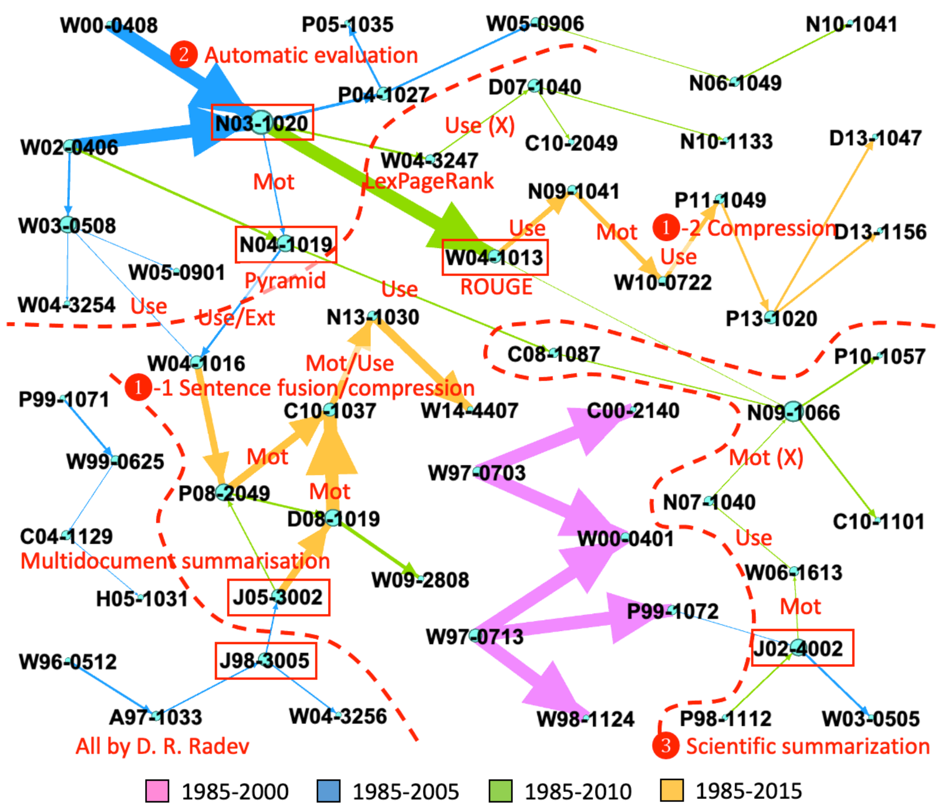

Figure 15. Main Path Network Extracted from `AANSum_plus_add_Use`.

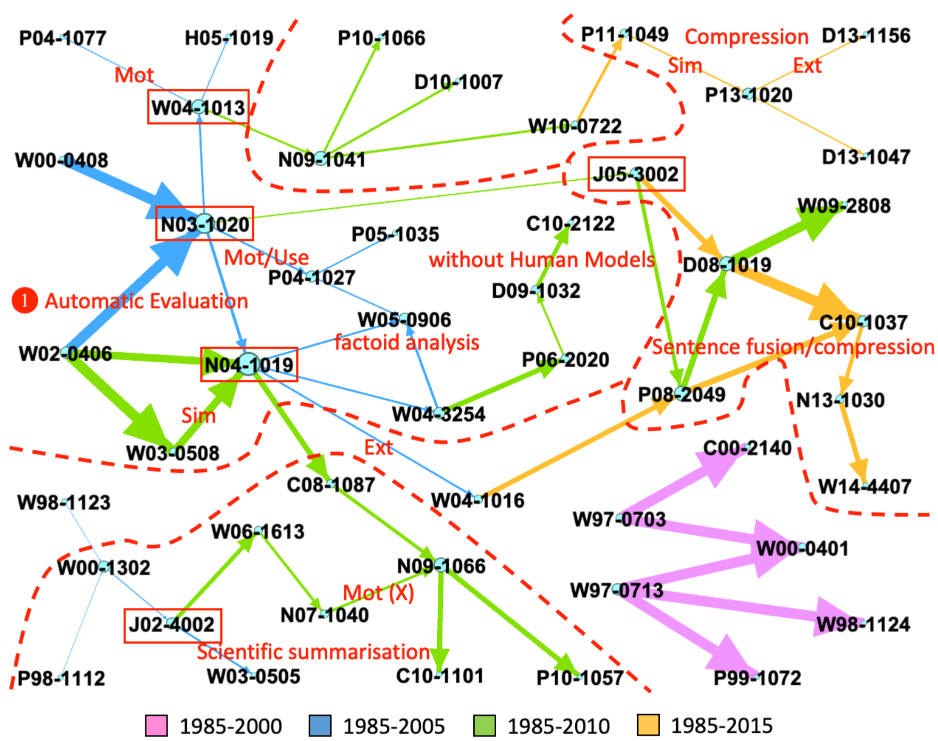

Figure 16. Main Path Network Extracted from `AANSum_plus_add_Sim`.

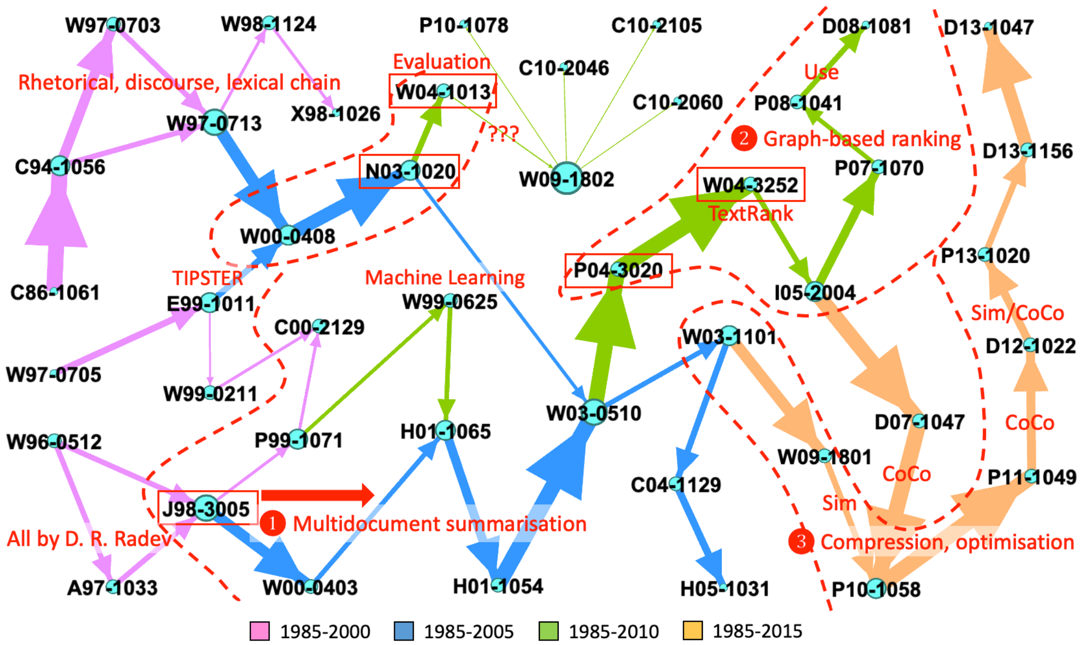

Figure 17. Main Path Network Extracted from `AANSum_del_Bkg_Fut`.

6. Quantitative Analysis

Few studies touched quantitative MPA evaluation. Filippin (2021) claimed that it is questionable if a main path is representative of the real technological trajectory because, based on domain experts' opinions, main path may be "limited to a much narrower neighborhood of the technology space than it really is" and may miss many crucial studies and big players of the analysed field. Huang et al. (2022) claimed to have achieved better convergence, which was only qualitatively justified. The current situation called us to propose a three-way framework for quantitative MPA evaluation. The first drawback pointed out by Filippin implies that a good main path should have a good *coverage* of the scientific topics of an analysed domain. It should also include as many critical studies as possible. We name this aspect the *pertinence* of main path. Furthermore, according to Huang et al., nearby main path nodes should exhibit a certain level of local clustering and show higher topical *coherence*. Our framework evaluated all these three aspects.

6.1. Topic Modelling

Coverage and coherence were both defined based on topic modelling, here LDA (Blei et al., 2003) trained using the Gensim package[11]. Each article u in the citation network, denoted as *CN*, was represented by its topic distribution $\mathbf{u} = [u_1, \ldots, u_t, \ldots, u_T]$, where T is topic number, u_t is the probability of article u belonging to topic t, and $\sum_{t=1}^{T} u_t = 1$. Two issues arose: the right value of T and the right number of training epochs P (to avoid overfitting LDA training). Supplementary Sect. G details how to decide these values. In summary, we trained several LDA models with a range of values of T for evaluation and reported the average. For `AANPar`, T values fell in {10, 11, …, 20, 22, 24, 26}. For `AANSum`, and `AANMT`, the maximum value of T was set to 20. The right value of P was set to 50, 40, and 50 for `AANPar`, `AANSum` and `AANMT` respectively.

6.2. Topical Coverage

Let *MN* denote an extracted MPN. *Topical coverage* measures how well *MN* covers the topics of the analysed domain. It is approximated by the closeness between the topic distribution of *MN*, denoted as $dist_{tpk}(MN)$, and the topic distribution of *CN*, denoted as $dist_{tpk}(CN)$, both of which are averaged over the enclosed publications. In evaluation, we used Hellinger distance to measure topical coverage, defined below:

$$cov_{tpk}(MN, CN) = D_{Hellinger}\left(dist_{tpk}(MN), dist_{tpk}(CN)\right), \tag{1}$$

[11] https://radimrehurek.com/gensim

where the Hellinger distance between two vectors **u** and **v** is defined as

$$D_{Hellinger}(\mathbf{u},\mathbf{v}) = \frac{1}{\sqrt{2}}\sqrt{\sum_i\left(\sqrt{u_i}-\sqrt{v_i}\right)^2}. \tag{2}$$

The smaller the Hellinger distance is, the better topical coverage is in our sense. Table 9 shows the results. Each "Δ%" column shows the difference of the corresponding semantic MPN from the vanilla MPN in percentage format. Thus, a positive percentage means a decrease in topical coverage and a negative percentage means increase. The upward and downward arrows signify a further increase and decrease from the semantic MPN in the column to the left. On all three datasets, compared to the semantics agnostic counterpart (the "`MPN`" column), topical coverage decreased (signified by upward arrows) by adding extension and motivation citations (the "`add_Ext_Mot`" column), but adding usage relations lead to improved topical coverage (signified by downward arrows in the "`plus_add_Use`" column). This is meaningful because publications linked with extension and motivation citations are technically closer. On the contrary, usage can be about a variety of different things, from algorithm and method to data and definition etc., and thus results in main paths that are topically more diverse. Two composite semantic MPNs were extracted: "`add_Combined`" corresponds to the composite semantic MPN which merged three semantic MPNs corresponding to "`add_Ext_Mot`", "`plus_add_Use`" and "`plus_Add_Sim`"; "`del_Combined`" corresponds to the composite semantic MPN which further merged the semantic MPN corresponding to "`del_Bkg_Fut`". The results proved that different types of semantic MPNs complemented each other and collectively worked better, i.e., covering and approximating the topic distribution of the underlying domain much better. Meanwhile, we also confess that better coverage was partially because composite semantic MPNs were larger in size (also see Table 11).

Table 9. Topical Coverage of Main Path Networks.

	`MPN`	`add_Ext_Mot`		`plus_add_Use`		`plus_add_Sim`		`add_Combined`		`del_Bkg_Fut`		`del_Combined`	
	cov_{tpk}	cov_{tpk}	Δ%	cov_{tpk}	Δ%	cov_{tpk}	Δ%	cov_{tpk}	Δ%	cov_{tpk}	Δ%	cov_{tpk}	Δ%
`AANSum`	0.0611	0.0647	+6.79% ↑	0.0591	–0.87% ↓	0.0679	+13.92% ↑	0.0509	–15.20% ↓	0.0630	+5.25%	0.0441	–26.53% ↓
`AANPar`	0.0582	0.0700	+25.13% ↑	0.0496	–8.40% ↓	0.0420	–24.34% ↓	0.0387	–29.62% ↓	0.0694	+21.57%	0.0380	–32.43% ↓
`AANMT`	0.0696	0.0794	+24.78% ↑	0.0617	–2.34% ↓	0.0697	+9.34% ↑	0.0621	–2.08% ↓	0.0619	–3.93%	0.0497	–20.18% ↓

6.3. Topical Coherence

A perfect definition of coherence does not exist. We tried to analyse coherence by adapting the coherence definition originally proposed to evaluate topic model quality (Newman et al., 2010, p. 102). Given a main path network *MN*, we defined *topical coherence* as the mean of distances between all pairs of main path nodes:

$$coh_{tpk}(MN) = mean\{D(u, v), \forall(u, v) \in MN\}, \tag{3}$$

where $D(u, v)$ is the distance between the topic distributions of u and v. Again, Hellinger distance defined in Eq. (2) was used.

Table 10 shows the results of topical coherence evaluation. From the “Evaluate on *MPN*’ rows, again, we observed that adding usage citations (the “`plus_add_Use`” column) lead to worse topical coherence compared to using extension and motivation citations (the “`add_Ext_Mot`” column). This corroborates with the evaluation results of topic coverage, adding usage citations may introduce more diversified topics, which increases topical coverage at the expense of decreasing topical coherence. Contrastively, adding similarity citations (the “`plus_add_Sim`” column) improved topical coherence. This may be because similarity in research goal or methodology often happens between topically closer studies. On all three datasets, better topical coherence was consistently obtained (i.e., with a negative Δ% value) except on “`plus_add_Use`”, which demonstrated that semantic MPN may exhibit better semantic coherence than the semantics-agnostic counterpart. For comparison purposes, the lower half of the table shows the results evaluated on *CN*[*MN*], the citation subnetwork induced from *MN* with a few more unimportant citations. The results met our anticipation to see worse topical coherence. This conforms to our initial conjecture that semantically important citations may help improve semantic coherence.

Table 10. Topical Coherence of Main Path Networks.

		MPN	add_Ext_Mot		plus_add_Use		plus_add_Sim		add_Combined		del_Bkg_Fut		del_Combined	
		coh_{tpk}	coh_{tpk}	Δ%	coh_{tpk}	Δ%	coh_{tpk}	Δ%	coh_{tpk}	Δ%	coh_{tpk}	Δ%	coh_{tpk}	Δ%
Evaluate on *MN*	AANSum	0.5518	0.5350	–3.18%	0.5456	–1.30% ↑	0.5428	–1.70% ↓	0.5423	–1.84% ↓	0.5505	–0.26%	0.5484	–0.67% ↓
	AANPar	0.4504	0.4448	–1.34%	0.4600	+2.14% ↑	0.4504	–0.05% ↓	0.4488	–0.40% ↓	0.4472	–0.71%	0.4484	–0.48% ↑
	AANMT	0.4327	0.4261	–1.41%	0.4394	+1.61% ↑	0.4138	–4.43% ↓	0.4246	–1.77% ↑	0.4299	–0.70%	0.4266	–1.38% ↓
Evaluate on *CN*[*MN*]	AANSum	0.5709	0.5736	+0.51%	0.5642	–1.25% ↓	0.5529	–3.17% ↓	0.5631	–1.39% ↑	0.5720	+0.31%	0.5698	–0.16% ↓
	AANPar	0.4748	0.4602	–3.00%	0.4878	+2.79% ↑	0.4791	+0.95% ↓	0.4730	–0.31% ↓	0.4718	–0.61%	0.4726	–0.40% ↑
	AANMT	0.4492	0.4529	+0.85%	0.4576	+1.96% ↑	0.4489	–0.02% ↓	0.4535	+1.02% ↑	0.4461	–0.70%	0.4545	+1.20% ↑

6.4. Pertinence

Ranking pertinence measures whether an extracted MPN effectively and efficiently represents the significant studies of a research field. To approximate expert evaluation, we built three gold standard sets following Jiang et al.’s approach (2019, p. 12). The three gold standard sets, named `GS-Par`, `GS-Sum` and `GS-MT`, each contains 99, 204 and 197 papers respectively[12]. Note that, some gold standards were not recoverable by the way we built citation networks (refer to Supplementary Sect. B

[12] They are available at: https://github.com/xiaoruijiang/scirank/tree/main/datasets/gold_standards/ACL. Note that, to construct `GS-Par`, we referred to Jiang et al.’s gold standard papers about computational linguistics/natural language (Jiang et al., 2019), and manually picked out the papers about natural language parsing technologies, because the surveys we were able to find could not cover the whole area of natural language processing.

about experimental setup), so evaluation was based on the total number of gold standards *recoverable* from the citation network. For `GS-Par`, `GS-Sum` and `GS-MT`, the sizes of recoverable gold standards were 78, 151, and 176 respectively.

Taking MPN as an unranked set of papers, pertinence could be evaluated using classical information retrieval evaluation measures. Table 11 summarises the results, where V represents MPN size, GS represents the number of matched gold standard papers, and ^GS represents the maximal number of gold standards in the corresponding citation network or semantic citation network, followed by precision, recall and F1 score. We observed that, although a single semantic MPN might not return more matches, the composite semantic MPNs achieved much better ranking performance. Comparing the "`add_Combined`" and "`del_Combined`" rows against the "`MPN`" row, the recalls of the former were more than doubled on `AANPar` and `AANSum`, and gained more than 65% relative increase on `AANMT`. Recall that, it is extremely important that as many crucial studies as possible are detected by MPA. At the same time, F1 scores were also largely improved except on `AANMT_add_Combined`. In addition, from the last three rows, we saw that "`add_Combined`" and "`del_Bkg_Fut`" results also complemented each other. The most extreme case was on `AANMT`: the sum of recalls of "`add_Combined`" and "`del_Bkg_Fut`" was only slightly larger than the recall of "`del_Combined`", implying that they returned drastically different subsets of gold standards. This justifies our claim that semantic MPNs may exhibit higher diversity to complement each other, and it would be better to merge them for a more comprehensive view. Finally, the recalls and F1 scores on all three datasets corroborate with the findings of Filippin (2021) about MPA's unsatisfactory recognition rate of the most significant studies. Although semantic MPA proved to improve ranking pertinence by a large margin, there seemed to still large space to improve recall. To achieve this, we guess that it may be helpful to start and guide main path exploration by first ranking and selecting important publications in some way (Baek et al., 2014; Zhang et al., 2014; Tao et al., 2017; Ding et al., 2021).

Table 11. Evaluation Results of Pertinence of Main Path Networks.

	GS-Sum/AANSum						GS-Par/AANPar						GS-MT/AANMT					
	^GS	V	GS	Prec.	Rec.	F1	^GS	V	GS	Prec.	Rec.	F1	^GS	V	GS	Prec.	Rec.	F1
MPN	**78**	44	15	34.09%	19.23%	24.59%	**151**	75	31	41.33%	20.53%	27.43%	**176**	90	29	32.22%	14.72%	20.21%
add_Ext_Mot	59	34	15	44.12%	19.23%	26.79%	118	80	26	32.50%	17.22%	22.51%	156	51	18	35.29%	9.14%	14.52%
plus_add_Use	66	54	20	37.04%	25.64%	30.30%	134	82	36	43.90%	23.84%	30.90%	170	67	20	29.85%	10.15%	15.15%
plus_add_Sim	69	48	19	39.58%	24.36%	30.16%	138	109	49	44.95%	32.45%	37.69%	171	62	19	30.65%	9.64%	14.67%
add_Combined	69	74	**24**	32.43%	**30.77%**	**31.58%**	138	162	**63**	38.89%	**41.72%**	**40.26%**	171	110	**29**	26.36%	**14.72%**	18.89%
del_Bkg_Fut	77	44	13	29.55%	16.67%	21.32%	151	68	30	44.12%	19.87%	27.40%	176	78	28	35.90%	14.21%	20.36%
del_Combined	77	99	**31**	31.31%	**39.74%**	**35.02%**	151	201	**75**	37.31%	**49.67%**	**42.61%**	176	168	**48**	28.57%	**24.37%**	**26.30%**

7. Conclusions

This paper advocated a novel semantic main path network approach for extracting the scientific backbone from a citation network based on citation function analysis. First, based on per-class performances analysis, the best models for extension, motivation, usage, similarity, neutral (equiv. background) and future work citation were picked from 55 contextualised classification models trained from 11 model architectures based on SciBERT. Then, four types of semantic citation networks were created by gradually adding extension and motivation citations, usage citations, similarity citations in a recall-oriented fashion, and by further deleting neutral and future work citations in a precision-oriented way. On each semantic citation network, semantic main path network was extracted by merging the top-*K* key-route main paths extracted from different time slices of the network. Meanwhile, for the first time, this paper performed quantitative main path analysis evaluation based a proposed a three-way framework consisting of topical coverage, topical coherence and ranking pertinence. The effectiveness of this approach was demonstrated on three computational linguistics subfields, namely natural language parsing, automatic text summarisation and machine translation.

Qualitative analysis showed that each semantic main path network was able to reveal novel topic branches, new important papers of existing branches, and the development pathways between papers and branches, thus provided complementary views of domain evolution. For example, for large domains such as natural language parsing that were guided by a few seminal studies (like Penn Treebank) and ground-breaking shared tasks, the semantic main path networks were much better at finding these representative works, such as the two early shared tasks on (multilingual) dependency parsing and more future shared tasks on a plethora of topics including semantic dependency parsing, semantic role labelling and dependency parsing of morphologically rich languages, most of which were missed by traditional main path analysis. For automatic text summarization, the semantic main path network approach was able to find an important novel branch about summarization evaluation and the branch about optimization methods for summarization, at the same time enrich the multi-document summarization, graph-based ranking and sentence fusion/compression branches that were recognised by the traditional approach.

Merging multiple semantic main path networks resulted in significantly better topical coverage. When main path analysis is seen as a method to return an unordered set of top-ranked studies, the composite semantic main path networks achieved much better ranking pertinence based on expert-selected gold standards, thus proved to be more comprehensive representations of scientific development. In addition, extension, motivation and similarity citations proved to achieve better semantic coherence on all three datasets than traditional approaches which ignore citation semantics, but adding usage citations may introduce topical diversity, which resulted in lower coherence but higher coverage. In the extracted semantic main path networks, most recognised citation relations were more relevant to uncovering the knowledge flow among scientific ideas. On the contrary, many main path papers were connected via incidental citations, such as neutral citations, using traditional approaches. Therefore, we conclude that the semantic main path network approach can discover more pertinent topic branches, uncover more coherent knowledge flows, and provide a more comprehensive scientific domain representation.

8. Acknowledgement

Xiaorui Jiang is partially supported by National Office for Philosophy and Social Sciences of China (18ZDA238).

9. REFERENCES

Abu-Jbara, A., Erza, J., & Radev, D. (2013). Purpose and Polarity of Citation: Towards NLP-based Bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'13*), 596–606. https://aclanthology.org/N13-1067

Bae, D.-H., Hwang, S.-M., Kim, S.-W., & Faloutsos, C. (2014). On Constructing Seminal Paper Genealogy. *IEEE Transactions on Cybernetics*, *44*(1), 54–65. https://doi.org/10.1109/TCYB.2013.2246565.

Batagelj, V. (2003). Efficient Algorithms for Citation Network Analysis. Available at: https://arxiv.org/abs/cs/0309023.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP'19)*, 3615–3620. https://aclanthology.org/D19-1371

Blei, D., Ng., A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022. https://www.jmlr.org/papers/v3/blei03a.html

Chen, L., Xu, S., Zhu, L, Zhang, J., Xu, H., & Yang, G. (2022). A semantic main path analysis method to identify multiple development trajectories. *Journal of Informetrics*, *12*, 101281. https://doi.org/10.1016/j.joi.2022.101281

Cohan, A., Ammar, W., van Zuylen, M., & Cady, F. (2019). Structural Scaffolds for Citation Intent Classification in Scientific Publications. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'19)*, 3856–3896. https://aclanthology.org/N19-1361

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41*(6): 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9

Dietz, L., Cickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning (ICML'07)*, 233–240. https://doi.org/10.1145/1273496.1273526

Ding, J., Xiang, T., Ou, Z., Zuo, W., Zhao, R., Lin, C., Zheng, Y., & Liu, B. (2021). Tell Me How to Survey: Literature Review Made Simple with Automatic Reading Path Generation. Available at: https://arxiv.org/abs/2110.06354

Dong, C., & Schäfer, U. (2011). Ensemble-style Self-training on Citation Classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, 623–631. https://aclanthology.org/I11-1070

Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2003). Why do we need algorithmic historiography? *Journal of the American Society for Information Science and Technology*, *54*(5), 400–412. https://doi.org/10.1002/asi.10226

Ghosal, T., Tiwary, P., Patton, R., & Stahl, C. (2022). Towards establishing a research lineage via identification of significant citations. *Quantitative Science Studies* (Advance publication). https://doi.org/10.1162/qss_a_00170

Hassan, S.-U., Safder, I., Akram, A., & Kamiran, F. (2018). A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. *Scientometrics*, *116*, 973–996. https://doi.org/10.1007/s11192-018-2767-x

Hernández-Alvarez, M., Gómez, J.M., & Martínez-Barco, P. (2017). Citation function, polarity and influence classification. *Natural Language Engineering*, *23*(4), 561–588. https://doi.org/10.1017/S1351324916000346

Huang, C.-H., Liu, J. S., Ho, M. H-C., & Chou, T.-C. (2022). Towards more convergent main paths: A relevance-based approach. *Journal of Informetrics*, *16*, 101317. https://doi.org/10.1016/j.joi.2022.101317.

Hummon, N. P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks*, 11(1), 39–63. https://doi.org/10.1016/0378-8733(89)90017-8

Iqbal, S., Hassan, S.-I., Aljohan, N.R., Alelyani, S., Nawaz, R., & Bornmann (2021). A decade of in-text citation analysis based on natural language processing and machine learning techniques: an overview of empirical studies. *Scientometrics*, 126, 6551–6599. https://doi.org/10.1007/s11192-021-04055-1

Jha, R., Abu-Jbara, A., Qazvinian, V., & Radev, D.R., (2017). NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, *23*(1), 93–130. https://doi.org/10.1017/S1351324915000443

Jiang, X., Zhu, X., & Chen, J. (2020). Main path analysis on cyclic citation networks. *Journal of the Association for Information Science and Technology, 71*(5), 578–595. https://doi.org/10.1002/asi.24258

Jochim, C., & Schütze, H. (2012). Towards a Generic and Flexible Citation Classifier Based on a Faceted Classification Scheme. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING'12)*, 1343–1358. https://aclanthology.org/C12-1082

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistic*, *6*, 391–406. https://doi.org/10.1162/tacl_a_00028

Kim, J., & Shin, J. (2018). Mapping extended technological trajectories: Integration of main paths, derivative paths, and technology junctures. *Scienctometrics*, *116*(3), 12–17. https://doi.org/10.1007/s11192-018-2834-3

Kim, M., Baek, I., & Song, X. (2018). Topic diffusion analysis of a weighted citation network in biomedical literature. *Journal of the Association for Information Science and Technology*, *69*(2), 329–342. https://doi.org/10.1002/asi.23960

Kim, E.H.J., Jeong, Y.K., Kim, Y.H., & Song, M. (2022). Exploring scientific trajectories of a large-scale dataset using topic-integrated path extraction. *Journal of Informetrics*, *16*(1), 101242. https://doi.org/10.1016/j.joi.2021.101242.

Kuan, C.-H. (2020). Regarding weight assignment algorithms of main path analysis and the conversion of arc weights to node weights. *Scientometrics*, *124*, 775–782. https://doi.org/10.1007/s11192-020-03468-8

Kuan, C.-H. (2022). Does main path analysis prefer longer paths? *Scientometrics*. https://doi.org/10.1007/s11192-022-04543-y

Kunnath, S.N., Herrmannova, D., Pride, D., & Knoth, P. (2022). A meta-analysis of semantic classification of citations. *Quantitative Science Studies*, *2*(4): 1170–1215. https://doi.org/10.1162/qss_a_00159

Lauscher, A., Brandon, K., Kuehl, B., Johnson, S., Jurgens, D., Cohan, A., & Lo, K. (2021). MULTICITE: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'22)*, 1875–1888. https://aclanthology.org/2022.naacl-main.137

Liu, J.S., & Lu, L.Y.Y. (2012). An integrated approach for main path analysis: The development of the Hirsch index as an example. *Journal of the American Society for Information Science and Technology*, *59*(12), 1948–1962. https://doi.org/10.1002/asi.21692

Liu, J.S., Chen, H.-H., Ho, M. H.-C., & Li, Y.-C. (2014). Citations with different levels of relevancy: Tracing the main paths of legal opinions. *Journal of the American Society of Information Science and Technology*, *65*(12), 2479–2488. https://doi.org/10.1002/asi.23135

Liu, J.S., & Kuan, C.-H. (2016). A new approach for main path analysis: Decay in knowledge diffusion. *Journal of the Association for Information Science and Technology*, *67*(2), 465–476. https://doi.org/10.1002/asi.23384

Liu, J.S., Lu, L.Y.Y., & Ho, M. H.-C. (2019). A few notes on main path analysis. *Scientometrics*, *119*, 379–391. https://doi.org/10.1007/s11192-019-03034-x

Liu, J.S., Lu, L.Y.Y., & Ho, M. H.-C. (2020). A note on choosing traversal counts in main path analysis. *Scientometrics*, *124*, 783–785. https://doi.org/10.1007/s11192-020-03469-7

Lu, W., Meng, R., & Liu, X. (2014). A Deep Scientific Literature Mining-Oriented Framework for Citation Content Annotation. Journal of Library Science in China, 40(214), 93–104. (in Chinese) https://doi.org/10.13530/j.cnki.jlis.140029

Lucio-Arias, D., & Leydesdorff, L. (2008). Main-path analysis and path-dependent transition in HistCite-based historiograms. *Journal of the American Society of Information Science and Technology*, *27*(1), 25–45. https://doi.org/10.1002/asi.20903

Lyu, D., Ruan, X., Xie, J., & Cheng, Y. (2021). The classification of citing motivations: a meta-synthesis. *Scientometrics*, *126*, 3243–3264. https://doi.org/10.1007/s11192-021-03908-z

Munkhdalai, T., Lalor, J., & Yu, H. (2016). Citation Analysis with Neural Attention Models. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI'16)*, 69–77. https://aclanthology.org/W16-6109.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'19)*, 100–108. https://aclanthology.org/N10-1012

Radev, D.R., Muthukrishnan, P., Qazvinian, V., & Abu-Jbara, A. (2013). The ACL anthology network corpus. *Language Resource and Evaluation*, *47*(4), 919–944. https://doi.org/10.1007/s10579-012-9211-2

Su, X., Prasad, A., Kan, M.-Y., & Sugiyama, K. (2019). Neural Multi-task Learning for Citation Function and Provenance. In *Proceedings of the 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL'19)*, 394–395. https://doi.org/10.1109/JCDL.2019.00122

Tao, S., Wang, X., Huang, W., Chen, W, Wang, T., & Lei, K. (2017). From Citation Network to Study Map: A Novel Model to Reorganize Academic Literatures. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW'17 Companion)*, 1225–1232. https://doi.org/10.1145/3041021.3053059

Teufel, S., Siddharthan, A., & Tidhar, D. (2006a). An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue (SIGdial'06)*, 80–87. https://aclanthology.org/W06-1312

Teufel, S., Siddharthan, A., & Tidhar, D. (2006b). Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, 103–110. https://aclanthology.org/W06-1613

Teufel, S. (2010). The Structure of Scientific Articles: Applications to Citation Indexing and Summarization. Centre for the Study of Language & Information.

Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying Meaningful Citations. In Proceedings of the Workshops of Scholarly Big Data: AI Perspectives, Challenges, and Ideas at the 29th AAAI Conference on Artificial Intelligence. https://allenai.org/data/meaningful-citations

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All You Need. In *Proceedings of the Thirty-first Conference on Neural Information Processing System (NIPS'17)*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems*, *10*(1), 93–115. https://doi.org/10.1142/S0219525907000945

Yu, D., & Pan, T. (2021). Tracing the main path of interdisciplinary research considering citation preference: A case from blockchain domain. *Journal of Informetrics*, *16*(2), 101136. https://doi.org/10.1016/j.joi.2021.101136.

Yu, D., & Sheng, L. (2021). Influence difference main path analysis: Evidence from DNA and blockchain domain citation networks. *Journal of Informetrics*, *15*(4), 101186. https://doi.org/10.1016/j.joi.2021.101186.

Zhang, H., Li, L., Li, T., & Wang, D. (2014). PatentDom: Analyzing Patent Relationships on Multi-View Patent Graphs. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM'14)*, 1369–1378. https://doi.org/10.1145/2661829.2662031