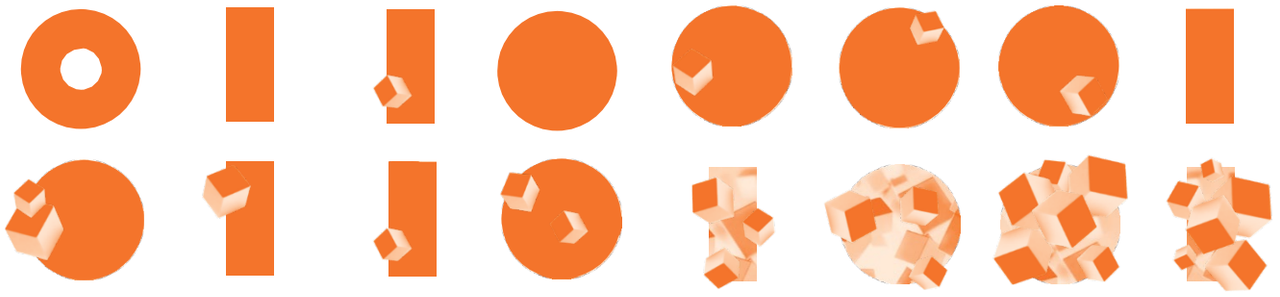

Technical Background and Methodology for the Elsevier's Artificial Intelligence Report

Mark Siebert, Curt Kohler, Anthony Scerri, Georgios Tsatsaronis
Elsevier

Abstract

Abstract—In this paper we give the background of the technical work conducted in the framework of preparing Elsevier's AI report. It explains how we are using a corpus-based approach and keyword extraction and seeding to focus on articles that potentially represent the core of the field. We then utilize both supervised and unsupervised machine learning approaches to filter and analyze this corpus with the aim to understand its trends, subfields, the connections between them, as well as how they evolve over time

Introduction

The field of Artificial Intelligence (AI) is broad, dynamic and rapidly evolving. Despite its rather young history with the term being coined in 1956 during a workshop in Dartmouth, it now encompasses a variety of sub fields, ranging from areas of generic content such as perception, learning or reasoning; to more application-specific content such as solving games, diagnosing diseases or real-time translation [1]. Common topics in AI, in turn, are attracting wide attention of researchers from different disciplines such as mathematics, linguistics and psychology; law- and policies makers as well as numerous industries. This nature and pace expound why the research landscape still lacks a united definition.

In this paper, we are describing an end-to-end methodology to identify research patterns of AI given a large-scale collection of scientific articles (see chapter "Approach and Sources").
First, we used expert knowledge and an in-house solution to extract and refine relevant concepts from selected sources, such as textbooks, expert panels, patents and public available sources as described in the chapter "A. Keyword extraction". Second, we use these search terms to retrieve a collection of scientific articles associated with AI from Scopus, an abstract and citation database [2]. We trained a classifier to reduce the size of this corpus to only high relevant articles, which we refer as AI corpus, as described in chapter "B – Optimization" Once the scope was defined, we used an ensemble of unsupervised clustering algorithms to learn latent topic structures within the corpus, as described in "C – Clustering". We believe that with this 'bottom-up' strategy we can represent the structure of the broad and dynamic AI domain. Many alternatives exist along this process and the "Discussion" suggests continuing research in this field.

Approach and sources

Different approaches exist to delineate a research field. They range from expert-selection of a publication set over expert-selection of keywords up to unsupervised approaches, like SciVal Topics, based on document citations. They vary in the influence of experts and depend on the availability and trust of (seed) information.

As no given set of publications as defined AI corpus exists, we followed good practice from the past to work with AI experts on a set of keywords, allowing to extract a reference corpus of documents. It became quickly apparent that other than in other fields experts were not able to come to a joint view on the keywords or to provide an agreeable initial set of (seed) of publications. From desk research of other AI reports we additionally observed the different scoping and perspectives on AI as a technical, application or business field and the difference over time of what is labelled as AI.

Those aspects led us to explore the field bottom-up and to identify keywords from content sources representative to their perspective and use expert guidance to validate the selection and outcome. Internal and external experts acknowledged the approach to build a comprehensive base and the outcome as a good and broad scope of the field of AI, covering the potentially different schools and fields of AI not just now, but also into AI history. Attempts to identify viable patterns in the keywords, using related publications failed as they again relied on individual expert opinion, e.g., assigning, which keyword is an algorithm and which an approach. Yet, the exploration shed deeper light on the keyword selection of the different perspectives, e.g. the Media focus on AI applications.

Due to the absence of an agreed AI ontology to help prune or structure the keyword list, we worked with the resulting raw set of 6 million extracted publications from the ~800 keywords with the goal to optimize this instead of the keywords by reducing the number of false positives. This brings the additional opportunity to identify the AI share in keywords or other derived concepts, such as SciVal topics, although the optimization of the publication set comes with further expert influence. To optimize the publication set, resulting from a Scopus extraction using the 800 keywords as queries, experts from Elsevier Labs developed a supervised classifier, based on a gold training set of 1;500 documents, manually annotated based on a mix of criteria. With this the publication set was reduced by 90%, satisfying expert validation of the resulting document corpus of 600; 000 as the core publication set scoping the field of AI.

This was used as base for the AI trend analytics as well as further structuring of the field to identify and specify research fields within AI, using two unsupervised clustering approaches to help validate each other.

Figure 1 summarizes the method and approach of the three phases.

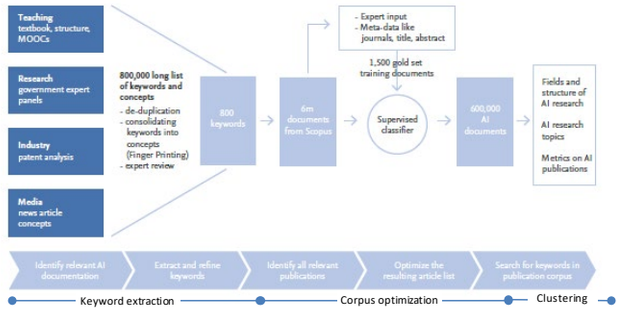

Fig. 1. Scoping and structuring approach for the field of Artificial Intelligence

The following Sections specify how we executed them in more detail. Goal of the approach was to offer and explore an expert-guided, bottom-up approach for an emerging and dynamic field, like AI, in contrast to top-down, expert-defined approaches of the past. Quantitative testing of the approach and results is suggested to be in focus for further research to help evolve bottom-up approaches and learn how to apply AI to define AI.

Perspec-tives	Sources	Key word approach	Key word results
Teaching	Two AI textbooks/-structure (Norvig structure [3]) (US, CHN), Nilsson [4] (Tsinghua), incl. Index). 50 AI course syllabi, incl. Coursera/edX, selected from leading Universities (Scival ranking by "AI" keyword). 10 Amazon books on "AI" keyword.	Extract content structure, abstracts, full-text (if Elsevier copyright). Manual key word extraction, cleaning and de-duplication. Fingerprint full-text into key words/concepts, based on general fingerprinting ontology. Fingerprinting, using full Scopus publication body, incl. conference papers and multiple cross-sector thesauri.	500k textbook concepts. Circa 2,500 concepts from books & courses.
Research	Suggestions from researchers in government expert panels (funding program design)	Expert-validated key word list on AI, based on Scopus analytics, including conference papers and multiple cross-sector thesauri.	Circa 50 key words – curated by experts.
Industry	Cooperative patent classification (CPC [5]) - patent codes G06N (Computer Systems based on specific computational models)	Extract CPC category structure and key words. Extract all patent titles, abstracts and full-text (circa 7,000 in the G06N category. Extract concepts with fingerprinting from the patent info. Fingerprinting, using full Scopus publication body, incl. conference papers and multiple cross-sector thesauri	Circa 100 key concepts from patent categories. Over 300,000 concepts from patent full text/abstracts.
Media	Annotated concepts and technologies; www.aitopics.org (official publication of the AAI on AI news)	Direct use of concept list	210 unique concepts

Fig. 2. Data sources and results for keyword extraction of AI perspectives

The first step was to create a set of keywords, refer to as search terms, associated as relevant to AI as possible. We use several content sources and text mining solutions alongside with selected domain experts to extract and refine these search terms as described in III-A. Consequently, we used these search terms to retrieve a raw collection of over 6

million scientific articles from Scopus, our in-house abstract and citation database [2]. This set has been reduced to a final list of approx. 600; 000 articles associated as relevant by means of a classifier described in III-B.

A. KEYWORD EXTRACTION

To extract bottom-up a set of keywords, referred to as search terms, we use several content sources and text mining solutions alongside with selected domain experts to extract and refine them. We selected with expert input content that represents each perspective (per Figure 2) and were accessible to us. We examined that content in various ways: text-mining bottom-up concepts from full-text (e.g. patents), using expert-curated lists of keywords (e.g. research), or referring to existing ontologies (e.g., aitopics.org), to obtain a long list of more than 800; 000 keywords, with mixed quality.

To clean and optimize the keyword list we used a number of filters and iterations.

- Exclude duplications, which were about 50
- Exclude general language words, like articles, verbs or stop words
- Exclude general terms
- Reduce term variations to a representative minimum and test sensitivity in Scopus
- Conceptualize it using fingerprints, which reduces connected terms to concepts
- Exclude clear non-AI or Computer Science terms, e.g. stemming from examples/cases in the documents
- Exclude specific, non-popular brands

For text-mining we relied on the Elsevier Fingerprint Engine. This identifies concepts and their importance in any given text by using a wide range of thesauri and data-driven controlled vocabularies covering all scientific disciplines, and by applying a variety of Natural Language Processing (NLP) techniques. The long list of keywords was reduced with this to circa 20,000 key phrases.

A first round of expert cleaning reduced this list further to 3,171 unique keywords, based on general relevance to Computer Science in a broad sense. In a second round of expert validation we excluded further terms from the list by reducing general Computer Science terms and focusing on AI relevance. This resulted in a unique keyword master list containing 797 terms, which we offered, together with the down-selection process for external expert comments.

Attempts to find an agreeable set of shared keywords, representative enough to scope the breath of the field and being at the same time specific enough to AI were not successful.

Neither one perspective stood out nor an overlapping kernel was strong or representative enough to delineate the field. Yet they brought interesting insights from analysing the overlap of keywords between the perspectives. We experimented with multiple approaches, such as semantic consolidation and simplification of similar terms, keyword co-occurrences, or using publication weight and expert structures to apply better classifications.

Optimizing the keyword list by further consolidating the terms by e.g. truncating them (e.g. bio*, multi-agent*, etc.) led to even more general terms and even higher rates of false positives. Additionally, the provenance of the data would have been lost to analyse which perspective originally provided the input for the term. An in-depth quantitative sensitivity analysis and testing was left for further research.

Optimizing the keyword list via keyword co-occurrences (using keyword combinations) made the corpus either too narrow or still full of false positives as it didn't reduce the keywords with broad or secondary meaning, such as predictive model, visual analytics, or classification tree. Yet keyword co-occurrence provided an interesting way of exploring sub-fields in AI.

Optimizing the keyword list by using corresponding publication size, publication development or weight and classifying the keyword frequencies according to initial expert input on AI types (e.g. algorithm, approaches, applications) and sub-structures (e.g., Learning, Decision making) felt again too biased and did not bring conclusive results on possible structures in the AI field. Yet they gave indications of the different focus of the perspectives, e.g., Media focussing on applications. Given the challenges to agree on an optimized list of keywords we focused on optimizing the resulting broad corpus of publications, using a supervised machine learning classifier.

B. OPTIMIZATION

The publication corpus was obtained by searching for each keyword in the titles, abstracts, and keywords of documents included in a Scopus May 2018 dataset, retrieving 5:7 million unique documents. Expert review of the corpus reveals that this contained a high number of false positives, caused by application terms (e.g. computer games), historical AI terms (e.g. finite elements), broad terms (e.g. ethical values), or other field terms (e.g. neural networks in biology).

A corpus of 6 million documents was too big to manually curate with an expert panel. Therefore, we focused on a training set for a supervised classifier, making use of machine learning technologies itself to scope the field of AI. Although further expert influence is introduced this way, external experts appreciated the better transparency into the corpus evolution and control, compared to a full unsupervised approach at this stage. Yet, unsupervised approaches were also suggested to further cluster and structure the field.

To develop the training set for the classifier, we randomly selected 10,000 document identifiers from the 5:7 million documents identified by the Scopus searches. We initially took 500 of these identifiers augmented with key textual fields from Scopus (abstract, title, keywords, etc.) and fed

them into a custom AWS Mechanical Turk [6] task where internal experts scored them as: Strongly AI, Moderately AI, or not AI.
The distribution of scores from this initial set indicated we would not generate enough positive AI training data, so we created a very simplistic screening classifier using the intuition being that documents that hit on more terms (with core and moderate terms more heavily weighted) were more likely to be AI documents. We ran all the 5,7 million documents through that screening classifier and assigned them to one of the three buckets. We then selected 1,000 random documents from the buckets as a second pool of potential training documents. From this second set of 1000 documents, we created two additional random sets of 500 documents and scored them as previously described. The sets were combined to create our 1, 500 document test set, using the following criteria:

- Number, percentage, and weighted value (initial expert score of strong AI) of query terms in a document
- Weighed value of all query terms using the length of the documents title and abstract
- Number and percentage of ASJC codes assigned to the document, AI ASJC and Computer Science ASJC codes assigned to the document
- Computer Science Subject Area code
- Presence of an Abstract

Testing different classifiers, like Decision Trees and Nave Base to help the classifier differentiate between AI and not-AI documents, a Random Forest model predicted the AI scope with circa 85% accuracy best. The entire set of 5,7 million documents was run through the model to generate predictions that were used to reduce the number of documents identified as AI Research to more than 600, 000 documents.
To help experts validate the results we mirrored the appearance of resulting publications back to each individual keyword. The keywords would on the one hand-side help to see the focus of the corpus (if e.g. keywords that are less of core AI appear high in the ranking of "AI-count") and would provide a fourth option to optimize the keyword list, if the corpus seemed wrong, e.g., cutting out further keywords in the base corpus using a certain threshold of AI publications per keyword.

Figure 3 shows the absolute number of AI and Non-AI publications from the 600,000-corpus associated with the keyword, resulting in a Non-AI share (%). This allows to rank the AI-relevant and non-AI-relevant ones. To prune the keyword-list one would need to define a threshold at which level AI-relevance to cut. Since all keywords yield some AI papers this approach already illustrates the challenge of losing AI papers. For instance, cutting off by 50%, non-AI would exclude a lot of the over 25; 000 AI-related papers around Robotics or 14; 000 papers from expert systems.

Especially in the middle (20 - 50%) of the ranked list, we can identify most of the AI application fields in the keywords, that are interesting to keep.
With the aim to provide a rather comprehensive view of AI and AI publications appearing in many fields and keywords, we decided to not further cut the keyword list, but to work with the resulting corpus. Experts suggested that the rate of false positives, as individual impressions from checking the corpus, is low enough to proceed with analytics. Yet, this test provided the opportunity to now illustrate the AI share in other keyword clusters, like SciVal Topics and brought the insight that societal relevant AI application fields range in an AI share of 20 - 50%. This might be an interesting starting point for further application-focused AI research.

	Initial Keyword	**AI count** *(used as analytics base)*	**Non-AI count**	**% non-AI**
High- ranked	Back-propagation Neural Network	8107	70	0.9%
	Back-propagation Algorithm	4029	36	0.9%
	Cohen-Grossberg Neural Networks	613	6	1.0%
	Genetics-based Machine Learning	170	2	1.2%
	Neural Networks learning	1390	19	1.3%
Mid-ranked	self-driving car	234	225	49.0%
	Automatic Translation	438	390	47.1%
	Soccer Robots	469	408	46.5%
	Personal Assistant Systems	12	10	45.5%
	Autonomous Mobile Robot	2115	1438	40.5%
Low- ranked	nonlocality	19	3768	99.5%
	Bias Currents	31	6149	99.5%
	Choice Experiment	26	5608	99.5%
	Boltzmann Equation	57	12460	99.5%
	Biosensing	161	38938	99.6%

Fig. 3. High / Mid-/ Low-ranked keywords by AI-share, applying 600k AI publication corpus to 797 keywords, covering 1998-2017

Testing the corpus against single AI keyword results in SciVal, we see similar growth trends, but differences in the content of the two samples, as illustrated by the two word-clouds in Figure 4 (single keyword) and 5 (report corpus).

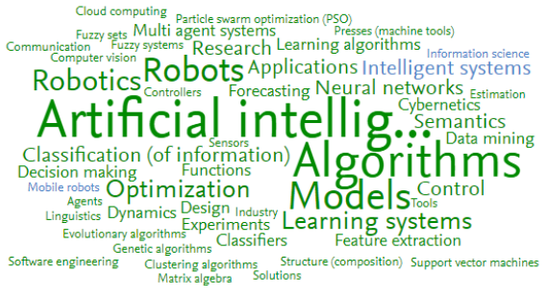

Fig. 4. Wordcloud, based on single keyword: Artificial Intelligence, SciVal, 2013-2017

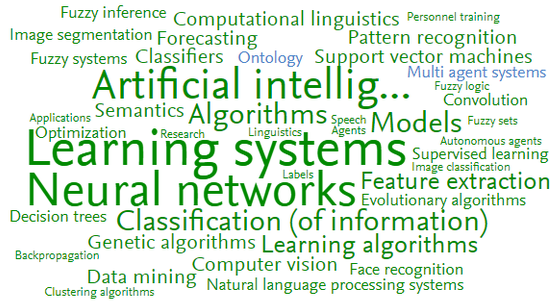

Fig. 5. Wordcloud, based on full AI report corpus, SciVal, 2013-2017

While the single keyword sample seems to focus more on general terms. Like AI, algorithms, models, is the report corpus highlighting specific approaches, like neural networks, classification or learning systems.

In any case, as Figure 6 illustrates, the report corpus covers significantly more subject fields. This underlines the initial intention to delineate the field broad, yet specific enough. It might provide a richer source, especially for analysis of the field structure, e.g. in terms of AI relevance in subject or application fields.

Subject Area Count Publication Year

Subject Area Sub-category Count

Entity	2013	2014	2015	2016	2017	Overall
AI_PublicationSet_1998-2017	299	301	297	302	300	323
Artificial Intelligence	78	83	83	98	104	119

Fig. 6. Comparison coverage subject fields between single keyword vs. AI report corpus, SciVal, 2013-2017

C. CLUSTERING

Learning more about the sub-structures of the AI field informs our report analytics and helps with deeper insights, e.g. which fields are the drivers of publication growth and how did the field evolve over time.

For this we refer to sub-structures within the now delineated field of AI, described by the publication corpus of 600;000 documents. Further research might investigate even broader approaches and explore fully unsupervised approaches for the field, such as other data sources, other seeds (selected publications) or testing different clustering algorithms and labelling.

Within the given field, we investigated and applied two clustering approaches: Louvain-clustering on keyword co-occurrences,and Bi-secting Kmeans clustering on the corpus itself. The first provides more supervision through the focus on the keywords, especially in terms of labelling. Other approaches, using stronger hierarchical information could also be considered, and we refer to future work for their application.

KEYWORD CO-OCCURRENCE CLUSTERING

We generated the matrix of co-occurrences between the keywords based on the document ID lists per keyword search. We then generated a record for each co-occurrence of a pair of terms for each document. This defined the global cooccurrence dataset that we then could filter based on core keywords, number of co-occurrences, or regions.

This resulting filtered dataset was then transformed into a network structure and processed via a Louvain clustering algorithm.

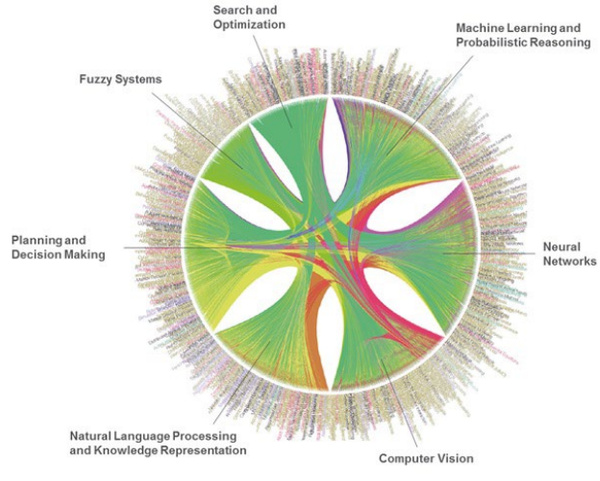

Fig. 7. Number of keyword co-occurrences based on at least 500 co-occurring documents, for the World, 2017; source: Scopus and Elsevier Fingerprint Engine

The Louvain algorithm attempts to merge individual nodes in the network together to clusters to maximize the number of

connections between the members of the cluster. The newly generated network of clusters then feeds into the next
iteration of the Louvain calculations. This continues until no further optimizations of the clustering can be made. Although the clusters appear separately, the co-occurrence points out the closeness of keywords and indicate that none of the clusters stands alone in the field.

Testing different visualization formats, like matrices or graphs, a Chord-Graph structure appeared most appealing to test users. Especially interactive functions, such as a time slicer or mouse-over received good resonance to facilitate the investigation of individual clusters or keywords, e.g. evolution of deep learning.
Figure 7 illustrates the resulting graph after experimenting with a series of tests with different publication thresholds and number of keywords (e.g., only core AI keywords from the classifier). We experimented with thresholds of 0, 100, 500 or 1;000 co-occurring documents. While 0 and 100 provide a rather full chart due to the multitude of cooccurrences, a threshold of 1;000 became to sparse and by region not representative anymore. The main clustering stayed stable across the thresholds with only smaller isolated clusters appearing/disappearing, probably due to similarity of keywords creating a nown cluster.
To illustrate the full field and first identify the clusters, we chose for all keyword despite the reduction in readability as a starting point. We kept a color coding for the provenance of the keywords to illustrate if there were clusters
predominantly provided or suggested by a specific group of AI practitioners. For deeper investigation and further research, e.g. on regional differences, by perspective, the threshold and number of keywords can be adapted and help increase readability. The report provides regional examples and further semantic analysis.

In summary, the discovered clusters in the field of AI are shown in Figure 5.

WORDEMBEDDINGS+KMEANS CLUSTERING

Neural network derived word embeddings are dense numerical representations of words that are computed by learning local co-occurrences. Their efficient training and ability to capture semantic and syntactical relatedness in various natural language processing (NLP) tasks including named entity recognition [7], part-of-speech tagging [8], and semantic role labelling [9], [10] of words have brought them much popularity. For this experiment, we trained our model on title and abstracts of approximately 70 million scientific articles from 30 thousand distinct sources such as journals and conferences. All articles are derived from Scopus abstract and citation database [2].

Cluster	Examples
Machine Learning and Probabilistic Reasoning	Machine learning, supervised learning, Nearest neighbor, Baysian networks
Neural networks	Deep learning, Backpropagation, Hopfield (and other) neural networks
Fuzzy systems	Fuzzy systems, Close-loop control
Search and Optimization	Swarm intelligence, evolutionary algorithms, genetic algorithms
Planning and Decision making	Distributed, adaptive systems, distributed artificial intelligence
Natural Language Processing and Knowledge representation	Hidden Markov, Word Sense Disambiguation, Natural Language Generation
Face recognition	Autonomous vehicles, Computer Vision, Image understanding, Robotics

Fig. 8. Summary Keyword co-occurrence clusters, World, 1998-2017; source: Scopus

After tokenizing, removal of stop-words and stemming the dataset contains a total of ca. 5.6 billion tokens (ca. 0.64 million unique tokens). Our word embeddings are obtained using a spark implementation [11] of the word2vec skip-gram model with hierarchical softmax as introduced by Mikolov et al [12], [13]. In this shallow neural network architecture, the word representations are the weights learned during a simple prediction task. To be precise, given a word, the training objective is to maximize the mean log-likelihood of its context. We have optimized model parameters by means of a word similarity task using external evaluation sets [14], [15], [10] and consequently used the best performing model.

This experiment resulted in five similar clusters to the initial 7 clusters of the co-occurrence clustering. Using wordcloud representations of the clusters identified that 2 initial clusters were “folded-in”, e.g., machine learning now appears as part of the other clusters. Given their strong co-occurrences this appears realistic and might be first indications for hierarchical relations, e.g. machine learning being an enabler for rather applied fields or other approaches, such as Computer Vision.

DISCUSSION

Along the scoping and definition process there are many alternatives ranging from different input data sets, through selection of seed documents up to alternative algorithms that all might influence the overall outcome. Key to this was to develop a bottom-up approach on larger scale data. It is an invitation to further testing and research on the path to a more robust and commonly accepted AI ontology.
Open to the whole process is the absence of an AI ontology and the missing use of hierarchical information, apart from some influence in the conceptualizing of keywords through our fingerprinting that makes use of other related (AI) taxonomies.
Other remaining challenges include clustering resolution, connectivity parameters, design of the graph, and labelling of the clusters as well as insights into the machine clustering.

ELSEVIER

REFERENCES

[1] Stuart Russell, Peter Norvig, and Artificial Intelligence. A modern approach. Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs, 25(27):79–80, 1995.
[2] Scopus search home page: www.scopus.com
[3] Russell S., Norvig P. (2010), Artificial Intelligence: A Modern Approach; (Third edition) - http://aima.cs.berkeley.edu/
[4] Nilsson, N. (1997), Artificial Intelligence: A New Synthesis - https://www.sciencedirect.com/science/book/9781558604674
[5] https://worldwide.espacenet.com/classification?locale=en_EP#!/CPC=G06N (US, EU, CHN); on CHN using CPC: https://www.epo.org/news-issues/news/2013/20130604.html
[6] AWS mechanical turk: https://docs.aws.amazon.com/mturk/index.html#lang/en_us : "Amazon Mechanical Turk is a web service that provides an on-demand, scalable, human workforce to complete jobs that humans can do better than computers, such as recognizing objects in photographs."
[7] Huy Do, Khoat Than, and Pierre Larmande. Evaluating named-entity recognition approaches in plant molecular biology. bioRxiv, page 360966, 2018.
[8] Cicero D Santos and Bianca Zadrozny. Learning character-level representations for part-of-speech tagging. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pages 1818–1826, 2014.
[9] Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. Jointly predicting predicates and arguments in neural semantic role labeling. arXiv preprint arXiv:1805.04787, 2018.
[10] Thang Luong, Richard Socher, and Christopher Manning. Better word representations with recursive neural networks for morphology. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pages 104–113, 2013.
[11] Spark home page: https://databricks.com/spark/about
[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.
[14] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: the concept revisited. ACM Trans. Inf. Syst., 20:116–131, 2001.
[15] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. Journal of Artificial Intelligence Research, 49:1–47, 2014.