Content Driven Enrichment of Formal Text using Concept Definitions and Applications

Abhinav Jain
IBM Research India
abhinavj@in.ibm.com

Nitin Gupta
IBM Research India
ngupta47@in.ibm.com

Shashank Mujumdar
IBM Research India
shamujum@in.ibm.com

Sameep Mehta
IBM Research India
sameepmehta@in.ibm.com

Rishi Madhok
Delhi Technological University
rishimadhok96@gmail.com

ABSTRACT

Formal text is objective, unambiguous and tends to have complex sentence construction intended to be understood by the target demographic. However, in the absence of domain knowledge it is imperative to define key concepts and their relationship in the text for correct interpretation for general readers. To address this, we propose a text enrichment framework that identifies the key concepts from input text, highlights definitions and fetches the definition from external data sources in case the concept is undefined. Beyond concept definitions, the system enriches the input text with concept applications and a pre-requisite concept graph that showcases the inter-dependency within the extracted concepts. While the problem of learning definition statements is attempted in literature, the task of learning application statements is novel. We manually annotated a dataset for training a deep learning network for identifying application statements in text. We quantitatively compared the results of both application and definition identification models with standard baselines. To validate the utility of the proposed framework for general readers, we report enrichment accuracy and show promising results.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; *Supervised learning by classification*; Neural networks; • **Applied computing** → *E-learning*;

KEYWORDS

Content Enrichment; Key Concepts; Concept Graph; Definition Extraction; Application Identification; Deep Learning

ACM Reference Format:
Abhinav Jain, Nitin Gupta, Shashank Mujumdar, Sameep Mehta, and Rishi Madhok. 2018. Content Driven Enrichment of Formal Text using Concept Definitions and Applications. In *HT '18: 29th ACM Conference on Hypertext and Social Media, July 9–12, 2018, Baltimore, MD, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3209542.3209566

HT '18, July 9–12, 2018, Baltimore, MD, USA

ACM ISBN 978-1-4503-5427-1/18/07…$15.00
https://doi.org/10.1145/3209542.3209566

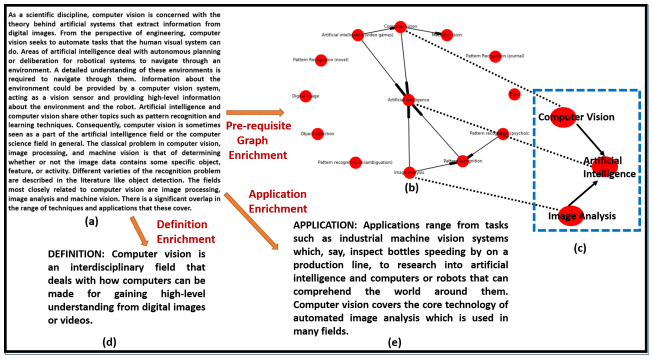

Figure 1: Output of enrichment system for formal text on Computer Vision which is undefined. (a) Input Text, (b)-(c) Pre-requisite graph, (d) Extracted Definition, and (e) Applications of key-concept "Computer Vision".

1 INTRODUCTION

Formal texts are characterized by coherency and completion, used to communicate knowledge. They contain technical terms which we call as key concepts. For example, consider an excerpt from a space and astronomy article *"What is dark matter?". "Dark matter may be made of baryonic or non-baryonic matter. To hold the elements of the universe together, dark matter must make up approximately 80 percent of its matter. Most scientists think that dark matter is composed of non-baryonic matter. The candidates for this are Neutralinos, massive hypothetical particles heavier and slower than neutrinos and sterile neutrinos."* The article is intriguing but filled with key concepts such as *"baryonic matter"*, *"neutralinos"* and *"sterile neutrinos"* for which descriptions are left out. With the exception of textbooks, formal text in daily usage (for example scientific articles, blogs etc.) lacks explanation of key concepts for the sake of brevity. In the absence of domain knowledge, it is difficult for readers to understand these key concepts within text and their relationships which leads to an incomplete semantic understanding of the presented text. We aim to solve this problem through a content enrichment system that analyzes the input text to conditionally enrich them with information in accordance with reader's discretion. We have sourced the information from Wikipedia. Wikipedia is a collaborative and open-source medium with reliable information on general topics referenced from verifiable and notable sources. Although we equip our enrichment framework by sourcing information from Wikipedia, we propose

an overall enrichment framework that can be easily extended to any available information source. For an average reader, we want to mitigate the cumbersome problem of searching the missing information through heaps of text via high quality augmentations in the form of definitions, real-life applications [6] and inter key-concept relationships that can provide more clarity to a key concept. By relationships, we mean scenarios where a concept X requires prior knowledge of the concept Y. The aforementioned augmentations are crucial for understanding a technical concept irrespective of reader's expertise in the corresponding domain.

In this work, we develop a framework for enrichment of formal text that consists of the following modules- (i) key-concepts extraction, (ii) definition identification, (iii) application identification, (iv) concept graph generation. The definition identification module checks for the presence of definitions for the extracted key-concepts and fetches the definitions of undefined key-concepts. The application identification module operates similarly to enrich the text with application statements. The concept graph generation module provides an overview of the pre-requisite relationships between extracted key-concepts to help the reader to understand the concept dependencies. The main contributions of our work are:

- We present a novel framework to conditionally enrich formal text with supplementary material that includes definitions, applications and concept graphs sourced from Wikipedia.
- Our method for definition and application identification utilizes LSTM networks and CNNs for sentence-level feature learning under a supervised setting.
- We created a (1) labeled dataset for learning application statements in formal text (2) datasets of formal text snippets obtained from *Scientific Articles* and lecture notes obtained from *MITOpenCourseware* to ascertain the effectiveness of our overall enrichment system.

2 RELATED WORK

One of the earliest approaches for enriching formal text, identified key-concepts and enriched the text with links to authoritative material [2], [1] found on Wikipedia. However, the semantics of the content and the text being enriched is ignored by both of these works. Another line of work has been explored in prior art to enrich formal text by providing a concept map. An approach is proposed in [8] that models concepts in vector space using their related concepts. An RefD score is proposed that measures the difference in the way the related concepts refer to each other. In [5], the method utilized cross-entropy and information flow separately to infer concept dependency relations. A method that jointly optimizes the two subproblems - key concept extraction and concept relationship identification for concept map extraction is proposed in [14].

Identifying definition statements from formal text has also attracted a lot of attention in the community. Some of the earliest approaches used hand-engineered features for their automatic extraction. For example, in [10], method proposed identifies star-patterns and word-class lattices from text for automatic definition extraction. A compendium of word level features for a weakly-supervised bootstrapping approach to classify sentences is proposed in [3]. To automate the learning of features for definition extraction, [7] models the problem as a supervised classification task, using LSTM to generate these features and outperforms non-Deep Learning based approaches.

The prior art partly addresses some of the challenges for enrichment of formal text. We combine some of these previous efforts and address some of their shortcomings to build a content-driven enrichment system for formal text. Specifically, we enrich the input formal text as seen in Fig. 1. In the next section, we present the details of the proposed framework.

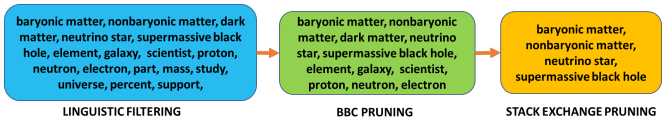

Figure 2: Key concept extraction output after every stage for the sample text shown in Table 1(a).

3 METHODOLOGY

Our overall methodology to enrich any input text consists of three phases. The initial phase extracts key concepts from the formal text, the second phase identifies the need for enrichment of these concepts, and the final phase conditionally enriches the input text with key concepts' definition and applications. This phase additionally generates a concept map from the input text which organizes key concepts based on their pre-requisite relationship with each other. We present the details of individual system components below.

3.1 Key Concepts Extraction

Extracting key concepts by exploiting training data limits the domain to work on. Thus in this section, we present a generic pipeline to reliably extract them from any input text.

Linguistic Filtering: The input text is first POS tagged using Stanford POS tagger [13]. Tagging is needed by linguistic filters that permit only specific strings for extraction without which strings such as *of the, is a* will also be extracted. We use following filters [2] - (i) P1 = C^*N, (ii) P2 = $(C^*NP)^?(C^*N)$ and (iii) P3 = A^*N^+ where N refers to a noun, P a preposition, A an adjective, and $C = A|N$.

Pruning: We leverage the word count dictionary of 90 million words BBC corpus [12] which is an up-to-date representation of general-science related vocabulary to identify stop-list of words such as good, day, voting, state, please, etc. It is further essential to filter candidate concepts such that they pertain to technical key concepts that may occur in formal text and require enrichment. In order to do so, we propose pruning using a corpus of such technical terms constructed using "tags" from StackExchange, SE. SE tags are used to annotate questions with a specific key concept that those questions pertain to. The SE Corpus approximately contains 60,000 tags from fields such as Mathematics, Physics, Electrical Engineering, Chemistry, Biology, Signal Processing etc. We illustrate the above stages in Fig. 2. Ultimately, we obtain a set of pruned key-concepts, $C = \{c_1, ..., c_N\}$, for which the need of enrichment is determined in the next phase.

3.2 Key Concept to Sentence Matching

Sentences quoting key-concept's definitions and applications have key-concepts as their subjects. Hence, we ensure unique association

Table 1: Excerpt from sample formal texts with annotation for some of the extracted key concepts (shown in bold).

Text		
Studies of other galaxies in the 1950s first indicated that the universe contained more matter than seen by the naked eye. Support for dark matter has grown, and although no solid direct evidence of dark matter has been detected, there have been strong possibilities in recent years. The familiar material of the universe, known as **baryonic matter**, is composed of protons, neutrons and electrons. Dark matter may be made of baryonic or **non-baryonic matter**. To hold the elements of the universe together, dark matter must make up approximately 80 percent of its matter. The missing matter could simply be more challenging to detect, made up of regular, baryonic matter. Potential candidates include dim brown dwarfs, white dwarfs and **neutrino stars**. **Supermassive black holes** could also be part of the difference.		
Key Concepts	**Definition**	**Application**
Baryonic matter	Yes	No
Supermassive black holes	No	No

i.e. one-to-one mapping between sentences and key-concepts for subsequent identifications. For example, *"The laws of thermodynamics define fundamental physical quantities (temperature, energy, and entropy) that characterize thermodynamic systems at thermal equilibrium."* has "*laws of thermodynamics*" as its subject since the sentences is its definition but not of key-concept "*thermal equilibrium*" which is merely mentioned in the sentence. For this, a set of sentences, S_i for every key-concept, c_i is created which contains all the sentences from the input text that have c_i as their subject using Stanford's Dependency Parser [4].

3.3 Application/Definition Identification

We formulate the identification phases as supervised binary classification problems. For every key-concept $c_i \in C$, we determine whether any sentence in S_i possesses a certain structure that marks the existence of concept's application or definition. Instead of hand-engineering these patterns, we employ Neural Networks to learn them from a carefully annotated dataset. We use CNN with LSTM because it excels at learning the spatial structure in input data. Our application and definition datasets have one-dimensional spatial structure in the sequence of words and the CNN should be able to pick out invariant features from the positive samples. These learned spatial features may then be learned as sequences by an LSTM layer. We have the following methodology executed on $\forall s_i \in S_i \forall c_i$:

(1) **Word Embeddings**: We encode Top-N frequent words as 300-dimensional GLOVE[11] vector embeddings.
(2) **Sentence Embeddings**: We add a one-dimensional CNN and max pooling layer after the Embedding layer which then feeds the consolidated features to the LSTM.
(3) **Classification**: Ultimately, LSTM feeds the learned sentence embedding to a dense network with logistic regression classifier which predicts labels of sentences in S_i. This overall learning is done on a carefully handcrafted dataset for which details are provided in Section 4.1.

3.4 Enrichment

After identification of key concepts, we mine Wikipedia's content to enrich the input text. We first provide the user with a pre-requisite relationship based concept map for better understanding of hierarchy present amongst identified key-concepts. Then we provide enrichment in terms of definitions and applications.

Pre-requisite Relationship Identification: We define the "pre-requisite structure" for a corpus as a graph, where nodes are key-concepts to comprehend, and a directed edge A → B corresponds to the assertion that "understanding A is a prerequisite to understanding". We identify the pre-request relationship between two key-concepts A and B by equally weighing the sum of the following similarity measures: (1) **RefD score** [8] using a threshold of $\theta = 0.02$ to determine the direction and existence of edge between A and B and (2) **Wikipedia link based Semantic Similarity** [15] to measure the semantic relatedness between A and B using the idea that if two concepts occur on the same page, they are more likely to be related to each other.

Enrichment with Definitions and Applications: For definitional enrichment, we deploy our definition and Application identification module on key-concept's Wikipedia page and identify those sentences which qualify as concept's definition and application.

Model	Prec(%)		Recall(%)		F1(%)	
	App^n	Def^n	App^n	Def^n	App^n	Def^n
LSTM	**88.09**	92.78	82.40	88.73	84.72	90.39
CNN	84.42	92.05	83.17	**92.79**	83.70	92.40
CNN-LSTM	87.21	**93.56**	**83.73**	92.25	**85.31**	**92.83**

Table 2: Performance of different models trained on Application and Definition Identification Model Training Datasets

4 RESULTS AND DISCUSSION

4.1 Tasks and Datasets

Definition Identification Model Training Dataset: We used the dataset provided by [9] to train our CNN-LSTM network for Definition Identification task. The dataset consists of 1,908 definitional sentences and 2,711 non definitional sentences created from Wikipedia which consists of domain-independent samples to prevent any kind of bias during learning. This makes the dataset apt for our purpose of enrichment of formal text because of its eligibility for domain-independent use.

Application Identification Model Training Dataset: Authors of the paper manually created and reviewed the annotated dataset from Wikipedia which consists of 3,000 positive candidates and 3,702 negative candidates for Application Identification task. We identified generic patterns within sentences which were classified as applications of key-concepts. Every positive candidate consists of (i) the key-concept being applied, (ii) a verb phrase showing how the key-concept is being applied and (iii) the field where the key-concept is being applied. Consider the following sentences:

(1) [*Scenery generators*]$_{concept}$ [are commonly used in] [*movies, animations and video games*]$_{fields}$.
(2) [*The COS cell lines*]$_{concept}$ [are often used by] [*biologists when studying the monkey virus SV40*]$_{field}$.
(3) [*In the production of semiconductor materials and devices,*]$_{field}$ [*octafluorocyclobutane*]$_{concept}$ [serves] as a deposition gas.

System Evaluation Dataset: To evaluate the overall effectiveness of our enrichment system, we created following Ground Truth datasets: (1) Lectured notes from MIT Open Courseware on 'Physics', 'Chemistry', 'Algebra' and 'Algorithms'. They contain a total of 80 educational texts (15 pages each and 10-15 key-concepts per page) and (2) 'Articles' dataset which consists of 100 articles(15-20 concepts each) from multitude of science magazines. Excerpt from some sample text is shown in Table 1. Lecture notes dataset have significant number of defined concepts in topics such as 'Probability', 'Chemical Reactions' etc but lacks their real-life applications. On the contrary "Articles" dataset is rich with formal texts targeted for specific demographics lacking definitions of many key-concepts.

		Ground Truth	
		Enrichment Required	Enrichment Not Required
Proposed Method	Enrichment Provided	TP	FN
	Enrichment Not Provided	FP	TN

Table 3: Notions for Enrichment Accuracy metrics

4.2 Evaluation Metrics

Model Learning: We evaluated the performance of all the learned models during 10-fold cross validation on definitions' and applications' dataset using Precision, Recall and F1-measure.

Enrichment System: To evaluate the performance of our enrichment system, we have the following metrics:

- **Key Concept Extraction (KCE)** : KCE, the first phase of our enrichment system is evaluated using using usual notions of Precision and Recall.
- **Overall Enrichment Accuracy (EA)**: Identification and Extraction phase is collectively evaluated using EA which is calculated using following notions of True/ False Positives/Negatives 3.

$$EA = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

4.3 Experiment Settings

Training : We trained the models using 10-fold cross validation. We restricted the Definition and Applications dataset to Top-1000 and Top-5000 frequent words respectively. GLOVE vector embeddings constituted the embedding layer. Following are the architectural details of different models stacked on top of the embedding layer:
(1) LSTM: LSTM (h=300 units) → Dropout, DL($p_{dropout} = 0.2$).
(2) CNN: Convolution Layer, CL(mask size=5, filter maps=128) → Max-Pooling Layer, ML(size=2) → CL(5, 64) → ML(2)
(3) CNN-LSTM: CL(5, 128) → ML(2) → LSTM(h=300) → DL(p=0.2).

The CNN layers used ReLU for activation. For training, we used Adam Optimizer to minimize log loss. A batch size of 32 was chosen, vanilla-LSTM model was trained for 3 epochs, CNN-model for 10 epochs and CNN-LSTM for 5 epochs.

Overall Testing : (a) Extract key-concepts, (b) Create $S_i \forall c_i$ using concept-sentence matching. (c) Run all identification models on $S_i \forall c_i$. (d) Obtain concept-dependency graph. (e) Based on (c), run identification model to mine relevant information from Wikipedia.

Pruning Phase	Prec(%)		Recall(%)		F1(%)	
	Edu	*Art*	*Edu*	*Art*	*Edu*	*Art*
LF	19.44	13.23	90.34	92.59	31.99	23.15
BBC Pruning	26.20	19.1	87.75	90.1	40.36	31.53
StackExchange	53.20	32.38	81.37	82.92	64.34	46.57
WikiRecall	78.68	72.04	76.35	81.71	77.50	76.57

Table 4: Performance of KCE phases in sequential order

4.4 Results

We report the results of all the KCE phases in Table4. Linguistic filtering does not have 100% recall because of maximal string matching. For example, 'isotherm' as a concept is ignored as it is part of the string 'validity of freundlich isotherm'. BBC pruning improves precision, but there is a small decrement in recall because of pruning of concepts like "Moore Voting" as it has a common(frequent) word 'Voting' in it. We observe that SE tags increases the precision due to the richness in terms of technical concepts they provide, however, we extract some additional unneeded concepts that may be technically sound but not relevant to the context of the text thus leading to increasing cognitive burden, redundancy and irrelevant data. This trade off however depends on how well the SE tag corpus is related to the context of the text. In Wikipedia based recall, we retain only those concepts which have corresponding Wikipedia articles. It is not part of our Key-concept extraction phase. However, we observe a decrease in recall due to unavailability of concepts' Wikipages. To showcase the effectiveness of SE pruning, we computed the results for Wikipedia based pruning without it: $Prec$: 75.81%, $Recall$: 75.46% and $F1$: 75.64%. From Table 2, it is quite evident from the results that the CNN-LSTM model preformed better as expected on both the datasets. Finally, we report EA for Definitions: 77.17% and 78.57% and Applications: 75.14% and 81.81% on Lectured Notes and Articles dataset respectively. For qualitative analysis, Fig. 1 represents the output of our enrichment system for formal text on "Computer Vision". The text neither contains its definition nor its applications, our system realizes this need and fetches them. Also, our system provides a concept graph to visualize the dependencies like understanding of "Computer Vision" is crucial before understanding "Artificial Intelligence".

5 CONCLUSION

We proposed a novel framework to enrich formal text with supplementary material. In the proposed approach, we extract key-concepts from text, identify the enrichment need using Deep Learning and finally enrich the text with definitions, applications and a pre-requisite concept graph to make the comprehension of the text easier. We also prepared the System Evaluation and Applications dataset. Lastly, we have done a quantitative analysis of the enrichment results to measure the effectiveness of our proposed framework. In future, we would like to validate the effectiveness of our enrichment framework on users with varying expertise by conducting a user study.

REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, and Krishnaram Kenthapadi. 2011. Identifying enrichment candidates in textbooks. In *WWW*. ACM.

[2] Rakesh Agrawal, Sreenivas Gollapudi, Krishnaram Kenthapadi, Nitish Srivastava, and Raja Velu. 2010. Enriching textbooks through data mining. In *DEV*. ACM.

[3] Luis Espinosa Anke, Horacio Saggion, and Francesco Ronzano. 2015. Weakly supervised definition extraction. In *RANLP*.

[4] Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*.

[5] Jonathan Gordon, Linhong Zhu, Aram Galstyan, Prem Natarajan, and Gully Burns. 2016. Modeling Concept Dependencies in a Scientific Corpus.. In *ACL*. ACM.

[6] Clyde Freeman Herreid. 1994. Case Studies in Science–A Novel Method of Science Education. *Journal of College Science Teaching* 23, 4 (1994).

[7] SiLiang Li, Bin Xu, and Tong Lee Chung. 2016. Definition Extraction with LSTM Recurrent Neural Networks. In *CCL*. Springer.

[8] Chen Liang, Zhaohui Wu, Wenyi Huang, and C Lee Giles. 2015. Measuring Prerequisite Relations Among Concepts.. In *EMNLP*. ACL.

[9] Roberto Navigli and Paola Velardi. 2010. Learning word-class lattices for definition and hypernym extraction. In *ACL*. ACM.

[10] Roberto Navigli, Paola Velardi, Juana María Ruiz-Martínez, et al. 2010. An Annotated Dataset for Extracting Definitions and Hypernyms from the Web.. In *LREC*.

[11] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. ACL.

[12] Tzipora Rakedzon, Elad Segev, Noam Chapnik, Roy Yosef, and Ayelet Baram-Tsabari. 2017. Automatic jargon identifier for scientists engaging with the public and science communication educators. *PloS one* 12, 8 (2017).

[13] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL*. ACL.

[14] Shuting Wang, Alexander Ororbia, Zhaohui Wu, Kyle Williams, Chen Liang, Bart Pursel, and C Lee Giles. 2016. Using prerequisites to extract concept maps fromtextbooks. In *CIKM*. ACM.

[15] Ian H Witten and David N Milne. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. (2008).