Guidelines for the Annotation of Scientific Concept Phrases

These guidelines discuss the annotation of scientific concept keyphrases in the context of the Open Research Knowledge Graph (ORKG) project. The goal of this work is to produce an annotated corpus to facilitate the development of natural language processing techniques to automatically extract keyphrases from scientific articles in order to aid populating the ORKG. To this end, we consider the core knowledge communicated in scientific literature as involving its attempted tasks, their objects of investigation, the materials used in the investigation, techniques employed, and the results obtained. A batch of manually annotated abstracts of scientific papers will serve as training and testing data for the development of the automated scientific information extraction NLP techniques.

We focus on how to recognise and annotate scientific concept keyphrases in scientific abstracts. Abstracts as general summaries of research investigations serve as an enriched resource of keyphrases as opposed to the full text article which would entail preselecting the main sections of the investigation to precisely extract its keyphrases.

A direct precursor to the attempt of annotating scientific keyphrases is the ScienceIE scientific information extraction task (Augenstein et al., 2017) which was the first community-wide initiative aimed towards the automatic extraction of scientific concepts at the keyphrase level from text. It facilitated the development of natural language processing techniques to automatically extract keyphrases from scientific articles from the following three types: (i) Process, (ii) Task, and (iii) Material, with an end goal of improved search of the content of scientific articles. The ScienceIE corpus comprised 500 paragraphs from full text articles annotated at the keyphrase level where the selected articles were uniformly distributed across three scientific domains, viz. Computer Science, Physics, and Material Science.

This work builds on the ScienceIE annotation initiative in two respects: 1) a finer granularity of scientific keyphrase annotation is attempted; and 2) 10 different STM (Scientific, Technical, and Medical) domains are considered. We attempt a finer granularity of annotations to allow a better semantic separation of concepts. However, we restrict the number of additional tags to balance better semantic separation with sufficient representation of the tags in our corpus.

The selected tags can be grouped into two main semantic categories, a primary tag group and a secondary tag group comprising one or more of the primary annotated concepts. The primary tag group comprises the following five semantic labels for keyphrases: (i) Task, (ii) Process, (iii) Method, (iv) Material, and (v) Data. The secondary tag group comprises the following two overlapping semantic labels for keyphrases: (i) Object and (ii) Result.

In the following text, we present our dataset of scientific abstracts and then give guidelines for tagging their scientific keyphrases.

Dataset

Our dataset is the Open Access Corpus of Scientific, Technical, and Medical Content provided by Elsevier. This corpus is a selection of 11 articles from 10 different STM domains that are the most published. The domains are agriculture, astronomy, biology, chemistry, computer science, earth science, engineering, materials science, math, and medicine.

In view of the practical feasibility of a manual annotation task in a multi-domain setting, we restrict ourselves to these 110 articles at this stage. However, in further development stages, the task can be attempted with other annotations or machine learning.

Scientific Keyphrase Concepts

The following concepts were chosen after a preliminary annotation of papers.

	Concept	Definition	Source
1	<Process>	Actions that alter or derive information from the state of the object(s) involved.	ScienceIE concept
2	<Task>	Includes smaller concrete research tasks (e.g. ‘powder processing’, ‘dependency parsing’) and broader research areas (e.g. ‘machine learning’).	ScienceIE concept
3	<Material>	Includes physical materials (e.g. ‘iron’, nanotube’), as well as datasets/corpora (e.g. ‘the ConLL-2003 NER corpus’).	ScienceIE concept
4	<Method>	A named process/model/specialized method is a Method	DEO
5	<Data>	The data themselves such as “velocity is 9m/s”, or quantified elements such as “three” in three epochs	DEO
6	<Object>	Entity/entities which is a product or main theme of an investigation	CoreSC (Liakata et al. 2010)
7	<Result>	Report of the specific findings of an investigation, given without discussion or conclusion being drawn	DEO

Annotation Procedure

1. For scientific keyphrases, we annotate definite noun phrases whenever possible.

2. We annotate quantifiable entities as Data.
3. We also annotate coreferring mentions across the abstract with the same type.
4. We annotate the exact material (including qualifiers) used for the experiment. For instance, “carbon atoms in graphene” is one Material in our annotations.
5. Confusion in tagging decisions is resolved using a tagging precedence where the tag appearing earlier in the list is preferred. The tag precedence is: Task > Method > Process > Data > Material.

Annotation Examples of Scientific Keyphrase Concepts

Primary Concept Tags

(i) Task

E.g.1] In the frame of the Land Use/Land Cover Area Frame Survey sampling of topsoil was carried out on around 22,000 points in 25 EU Member States in 2009.

“the Land Use/Land Cover Area Frame Survey” is annotated as <Task> since it is a broader research area

E.g.2] The path of the chirality delivery in the crystalline and chiral nucleotide-Co(II) complex has been studied based on X-ray single crystal diffraction analysis, liquid- and solid-state circular dichroism (CD) spectroscopy.

“X-ray single crystal diffraction analysis” and “liquid- and solid-state circular dichroism (CD) spectroscopy” are each annotated as <Task>

E.g.3] In this research, we developed a robust two-layer classifier that can accurately classify normal hearing (NH) from hearing impaired (HI) infants with congenital sensori-neural hearing loss (SNHL) based on their Magnetic Resonance (MR) images.

“classify normal hearing (NH) from hearing impaired (HI) infants with congenital sensori-neural hearing loss (SNHL)” is annotated as <Task>

Heuristics for identifying <Task> candidates

a. We identify <Task> in contexts such as the following: “Research into…”, “the recent progress in…”, “further enhancements in…”, “enabling …”, “[tool] for …”, “approaches to …”, “applied to …”, “[Material] for …”, “[method] is widely used for …”, “several works are dedicated to …”, “In the present work we use [Method] for …”, “We show that the method can be applied to …”, “Special attention is paid to …”, “used to perform …”, “a key challenge of …”

b. We annotate as <Task>, the phrases including the word “problem” or “studies.”

(ii) Process

E.g. 1) The transfer of a laboratory process into a manufacturing facility is one of the most critical steps required for the large scale production of cell-based therapy products.

“The transfer”, “a laboratory process”, and “the large scale production” are each annotated as <Process>

E.g. 2) The transterminator ion flow in the Venusian ionosphere is <u>observed</u> at solar minimum for the first time.

“The transterminator ion flow” and “solar minimum” are annotated are <Process>

The verb “observed” is not annotated as <Process> since it doesn’t act upon another object.

E.g. 3) It is suggested that this ion flow contributes to maintaining the nightside ionosphere.

“this ion flow” and “maintaining” are annotated as <Process>

E.g. 4) Modified protocols were developed for the automated system.

“Modified protocols” is annotated as Process.

The verb “developed” is not annotated as <Process> since it does not act upon another object.

E.g. 5) The management of cells aggregates (clumps) was identified as the critical step.

“The management” is annotated as <Process>

The verb “identified” is not annotated as <Process> since it doesn’t act upon another object.

E.g. 6) Cellular morphology, pluripotency gene expression and differentiation into the three germ layers have been used compare the outcomes of manual and automated processes.

“pluripotency gene expression”, “differentiation”, “compare”, and “manual and automated processes” are each annotated as <Process>.

Heuristics for identifying <Process> candidates

a. Verbs (e.g., measured), verb phrases (e.g., integrating results), or noun phrases (e.g., an assessment, future changes, this transport process, the transfer) are candidates for <Process>.
b. <Process> can be one of two things, an occurrence natural to the state/properties of the entity or an action performed by the investigators. In the latter case, however, it is a <Method> when the action is a named instance.

(iii) Method

E.g. 1) Here finite-element modelling has demonstrated that once one silica nanoparticle debonds then its nearest neighbours are shielded from the applied stress field, and hence may not debond.

“finite-element modelling” is annotated as a <Method>

Heuristics for identifying <Method> candidates

a. We annotate as <Method>, the phrases containing any of following words: simulation, method, algorithm, scheme, technique, system, function, derivative, proportion, strategy, solver, experiment, test, computation, program.

(iv) Material

E.g. 1) Based on the results of the LUCAS topsoil survey we performed an assessment of plant available P status of European croplands.

“European croplands” is annotated as <Material>

E.g. 2) The transfer of a laboratory process into a manufacturing facility is one of the most critical steps required for the large scale production of cell-based therapy products.

“a manufacturing facility” and “cell-based therapy products” are annotated as <Material>

E.g. 3) Cellular morphology, pluripotency gene expression and differentiation into the three germ layers have been used compare the outcomes of manual and automated processes.

“the three germ layers” is annotated as <Material>

(v) Data

E.g. 1) Based on the results of the LUCAS topsoil survey we performed an assessment of plant available P status of European croplands.

“the results” and “plant available P status” are annotated as <Data>

E.g. 2) Our analysis shows a status of a baseline period of the years 2009 and 2012, while a repeated LUCAS topsoil survey can be a useful tool to monitor future changes of nutrient levels, including P in soils of the EU.

“a status of a baseline period”, “nutrient levels”, and “P” are annotated as <Data>

E.g. 3) Observations near the terminator of the energies of ions of ionospheric origin showed asymmetry between the noon and midnight sectors, which indicated an antisunward ion flow with a velocity of (2.5±1.5)kms-1.

“asymmetry between the noon and midnight sectors”, “a velocity”, and “(2.5±1.5)kms-1” are annotated as <Data>

E.g. 4) “We established [a P fertilizer need map] based on integrating results from the two systems.”

“a P fertilizer need map” is annotated as <Data> overriding “a P fertilizer” as <Material> by the tag precedence annotation guideline.

Secondary Concept Tags

(i) Object

E.g. 1) Besides other basic soil properties soil phosphorus (P) content of the samples were also measured in a single laboratory in both years.

“soil phosphorus (P) content” is annotated with primary tag <Data> and “the samples” is annotated with primary tag <Material> which are both also the main objects of the investigation. Hence, they are additionally annotated with the secondary concept tag <Object>.

(ii) Result

E.g. 1) Meanwhile we found disparities of calculated input need and reported fertilizer statistics both on local (country) scale and EU level.

“calculated input need” and “reported fertilizer statistics” are annotated with primary tag <Data> which are part of the phrase “disparities of calculated input need and reported fertilizer statistics” annotated as <Result> since it is a result of the investigation.

References

Augenstein, Isabelle, et al. "Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications." *arXiv preprint arXiv:1704.02853* (2017).

Liakata, Maria, et al. "Corpora for the conceptualisation and zoning of scientific papers." (2010)