Doing Resarch in Natural Language Processing

Session 15: Scientific Writing: Proposals

Frank Keller

8 November 2023

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

Overview

Heilmeier's Questions

Things Heilmeier Doesn't Cover

Reading: Alley (2018), ch. 9 (pp. 174–198).

Please also look at Alley's web site, which has a lot of videos and additional materials: https://www.craftofscientificwriting.com/

Overview

Most scientific documents report completed work. But proposals describe work you are *planning to do.* They are required in a range of contexts:

- planning a student project (e.g., IP proposal, PhD proposal)
- applying for research funding (grant proposal, fellowship proposal)
- applying for an academic position (often requires a research statement)
- planning a research or development project for a company
- pitching a business idea to an investor

Many of the things we've learned so far readily apply to proposals.

Overview

Audience: Who are they, why are they reading, what do they know? For proposals, often a mix of expert reviewers and generalists (panel members, management).

Purpose and occasion: get your project funded/approved, so the focus is on persuasive writing.

Structure: not unlike a paper; but often detailed instructions or even a template is provided by the funding agency etc.

Precision and clarity, style, use of figures and tables: essentially the same advice applies as for other scientific writing.

Alley has a lot of good advice on proposals and lots of examples: how to write the intro, how to use of illustrations, etc.

Heilmeier's Questions

Heilmeier's Questions

George Heilmeier was a director of DARPA, back in the days when the agency funded high risk/high gain research. He developed a set of questions that every proposal should answer:

1. What are you trying to do? Articulate your objectives using absolutely no jargon.
2. How is it done today, and what are the limits of current practice?
3. What is new in your approach and why do you think it will be successful?
4. Who cares?
5. If you are successful, what difference will it make?
6. What are the risks?
7. How much will it cost? How long will it take?
8. What are the mid-term and final "exams" to check for success?

This is known as the *Heilmeier catechism.* Alley's advice complements this.

How to use Heilmeier's Questions

- Think about how each question relates to the proposed project.
- Write down an answer to each question: in full sentences, explicitly and specific
- Consider Q7: this determines the scope of your project (being overambitious is a common mistake!)
- If you've been given instructions, a fixed structure, a call for proposals, or a template: map that onto your answers
- Now make an outline with the required structure; make sure all of Heilmeier's questions are covered
- Turn the outline into a full text proposal

Bear in mind reviewers/markers will also be given a fixed structure when they assess your proposal! *Make their job easy:* show them where the relevant information is.

1. What are you trying to do?

Give an overview of the proposal, without jargon and acronyms:

- state the overall goal of your research
- break it down into distinct research questions/hypotheses
- list the research objectives required answer these questions:
  - what are the intermediate steps?
  - which methods will be used?
  - which data will be collected?
  - how will you evaluate?

The answer to Q1 should be in the summary and in the intro of the proposal.

2. How is it done today, and what are the limits of current practice?

This is the related work or background section:

- describe the current state of the art
- identify gaps in the literature, unanswered questions, limitations of current methods
- explain how you will fill these gaps
- now you're allowed to use jargon ...

3. What is new in your approach and why do you think it will be successful?

Most funders put a high premium on novelty:

- what is new in your proposal, for instance:
  - theory
  - approach
  - methods
  - datasets
- why is your approach not only novel but also *better:* higher accuracy, faster, less memory/compute, etc.
- your approach is not only novel and better, but *it will work*
- ideally, refer to pilot data, preliminary experiments, similar work you have done
- also talk about your plan B: what will you do if your initial idea doesn't work?

4. Who cares?

- Researchers in your narrow field will of course care
- But who else will be interested? Are there:
  - application to new problems in the same field
  - application in other disciplines
  - users in industry
  - benefits to the economy or society as a whole
- How will you make sure that your work reaches other researchers, disciplines, users in industry or society?

Funding agencies call this *relevance to beneficiaries,* and take it rather seriously.

5. If you are successful, what difference will it make?

Argue that your project will:

- create new knowledge
- solve important scientific, technological, or societal problem
- contributes to research infrastructure and technical capabilities
- benefit economic development
- train the next generation of scientists and engineers

Funding agencies call this *impact.* Again, taken very seriously. Why should the taxpayer pay for this project?

6. What are the risks?

In Q4 and Q5 you argued that your project has a lot of benefits and high impact. But what are the *risks?*

- all novel ideas carry risk, in science things often don't work as expected
- it is important to *balance* risk and reward
- propose some things that are safe, and some things that are risky, but come with a high pay-off
- give the impression that you have thought about the risk
- suggest ways of managing risk; have a plan B if things don't work

7. How much will it cost? How long will it take?

These questions help you figure out the scope of the proposed project:

- Can what you suggest be done by one person in one semester (IP) or in three years (PhD); by a team in five years (large project)?
- Think not only in terms of money, but also in terms of time and resources (infrastructure, expertise, data, compute)
- Often the call for proposals specifies a limit to how much time your have and how much budget you can ask for
- Often a timeline is required or a breakdown in terms of work packages (with person months for each)
- At this point, you often seen a Gantt chart in the proposal ...

8. What are the mid-term and final “exams” to check for success?

Explain how you will know that you’re on track:

- Most projects require a mid-term and final evaluation
- The IP requires an interim report, a final report, and a presentation
- The PhD requires an annual report and a final report (aka thesis), as well as an oral exam
- In your proposal, say how you will evaluate progress, how you will deal with lack of progress
- This is a good place to talk about evaluation methods, metrics, test sets

Over to You

Exercise 1

Sadly, we don't have time to study full proposals. But on the next pages you will find the summaries of three ERC proposals:

- Can you find answers to some of the Heilmeier questions?
- Which questions are not covered?
- Do the summaries contain additional information not requested by Heilmeier?
- Comment on the level of detail and technicality of the summaries.

Induction of Broad-Coverage Semantic Parsers (Titov)

In the last one or two decades, language technology has achieved a number of important successes, for example, producing functional machine translation systems and beating humans in quiz games. The key bottleneck which prevents further progress in these and many other natural language processing (NLP) applications (e.g., text summarization, information retrieval, opinion mining, dialog and tutoring systems) is the lack of accurate methods for producing meaning representations of texts. Accurately predicting such meaning representations on an open domain with an automatic parser is a challenging and unsolved problem, primarily because of language variability and ambiguity. The reason for the unsatisfactory performance is reliance on supervised learning (learning from annotated resources), with the amounts of annotation required for accurate open-domain parsing exceeding what is practically feasible. Moreover, representations defined in these resources typically do not provide abstractions suitable for reasoning. In this project, we will induce semantic representations from large amounts of unannotated data (i.e. text which has not been labeled by humans) while guided by information contained in human-annotated data and other forms of linguistic knowledge. This will allow us to scale our approach to many domains and across languages. We will specialize meaning representations for reasoning by modeling relations (e.g., facts) appearing across sentences in texts (document-level modeling), across different texts, and across texts and knowledge bases. Learning to predict this linked data is closely related to learning to reason, including learning the notions of semantic equivalence and entailment. We will jointly induce semantic parsers (e.g., log-linear feature-rich models) and reasoning models (latent factor models) relying on this data, thus, ensuring that the semantic representations are informative for applications requiring reasoning.

The Evolution of Linguistic Complexity (Smith)

Human language is unique among the communication systems of the natural world, providing our species with an incredibly flexible and powerful open-ended system of communication. This expressive power is underpinned by linguistic structure: we construct complex meaning-bearing utterances according to a set of rules and regularities which are conventionalised among speakers of a language. In my previous work I have shown that these fundamental structural features of language can be explained as a consequence of cultural evolution: structure evolves gradually as language is passed down through generations via learning and shaped by its repeated use for communication, in a process known as iterated learning. However, existing modelling and experimental treatments of iterated learning are limited in that they focus on the evolution of simple languages which permit expression of a relatively small and fixed set of concepts. Real human languages are enormously complex, both in the expressive power they afford and the rich and complex set of structural devices they provide for conveying meaning. In this project I seek to address this major outstanding question in evolutionary linguistics: why is language complex? I will tackle this daunting question by exploring two subsidiary questions: when and how does linguistic complexity facilitate acquisition, and how do expressive power and linguistic complexity evolve as a result of language transmission and use? Answering these questions will require an ambitious programme of modelling and experimental work, covering acquisition in individual adults and children, language use in interaction, and language evolution in populations. I seek to substantially advance our understanding of the cultural evolution of language by exploring how learning, expressive pressures on language use, and social complexity drive the evolution of linguistic complexity.

Translating from Multiple Modalities into Text (Lapata)

Recent years have witnessed the development of a wide range of computational methods that process and generate natural language text. Many of these have become familiar to mainstream computer users such as tools that retrieve documents matching a query, perform sentiment analysis, and translate between languages. Systems like Google Translate can instantly translate between any pair of over fifty human languages allowing users to read web content that wouldn't have otherwise been available. The accessibility of the web could be further enhanced with applications that translate within the same language, between different modalities, or different data formats. There are currently no standard tools for simplifying language, e.g., for low-literacy readers or second language learners. The web is rife with non-linguistic data (e.g., databases, images, source code) that cannot be searched since most retrieval tools operate over textual data. In this project we maintain that in order to render electronic data more accessible to individuals and computers alike, new types of models need to be developed. Our proposal is to provide a unified framework for translating from comparable corpora, i.e., collections consisting of data in the same or different modalities that address the same topic without being direct translations of each other. We will develop general and scalable models that can solve different translation tasks and learn the necessary intermediate representations of the units involved in an unsupervised manner without extensive feature engineering. Thanks to recent advances in deep learning, we will induce representations for different modalities, their interactions, and correspondence to natural language. Beyond addressing a fundamental aspect of the translation problem, the proposed research will lead to novel internet-based applications that simplify and summarize text, produce documentation for source code, and meaningful descriptions for images.

Things Heilmeier Doesn't Cover

Why You?

Most funding agencies require a *track record* section:

- say why you are well qualified to carry out the proposed work
- explain relevant previous work you have done; emphasize your successes
- describe your expertise in the required theory, experimental methods, evaluation techniques, etc.
- if the work will be done by a team explain why each team member is vital, and how they will contribute
- a description of the institution and its facilities and resources may also be required

This is a bit like a CV or a cover letter in a job application!

Why Now?

This is sometimes called *timeliness:*

- Why is the time ripe for the proposed project? Why should it be funded now rather than next year?
- Reasons could include:
  - new theoretical foundations make the work possible
  - new data, new results make it a logical next step
  - there's an urgent societal or economic need
  - disciplines have converged to make the proposed work feasible
- or it could also just be that the topic is currently fashionable (though best not so say it like that . . .)

Methods and Feasibility

Heilmeier's questions don't explicitly address methods and feasibility. But reviewers pay *a lot* of attention to this, and it's an explicit review criterion for most funders:

- describe how the methods follow from your research questions and objectives
- contrast them with alternatives, explain why your methods are more suitable
- describe in technical terms how you will achieve your objectives: which theories, algorithms, datasets, experimental paradigms will you use, how you will evaluate results, analyze data, etc.
- you want to convince the reader that the proposed work is *feasible,* and that you can achieve your objectives

This part of the proposal should be technical (even though other parts should be accessible to non-experts).

References

Alley, Michael. 2018. *The Craft of Scientific Writing*. Springer, New York, NY, 4 edition.