University of Windsor
Faculty of Science
School of Computer Science

COMP-8730
Natural Language Processing & Understanding
Winter 2024

#	Title	Due Date	Grade Release Date
2	**Literature Review (Related Work)**	Feb. 04, AoE	Feb. 12

This course is *research-oriented* and *project-driven* in which a research project should be defined and completed in the field of NLP within one semester. The objectives of the research project are to provide graduate students with:

- An experience with research procedure, in general, and research in NLP, in particular,
- Hands-on experience with NLP,
- Advancing state of the art in NLP while passing a grad course,
- An opportunity to present a research outcome at an international Computer Science conference,
- An opportunity to meet with scholars in the NLP community.

In the research project, we propose a solution(s) to a problem by implementing an algorithm like a software project. However, there are differences in some respects. For instance, while a software project may implement an existing algorithm, a research project should propose and implement a *new* algorithm that improves or addresses a particular aspect of a problem that the current algorithms overlook. Roughly, a research project has the following milestones (phases):

1) Proposal
2) **Literature Review**
3) Proposed Method (Formal + Code)
4) Experiment (Evaluation)
5) Presentation (Paper + Talk)

In this course, a manual is prepared to guide the students through each milestone. The current manual is for the second milestone: Literature Review (aka. Related Work), which further has the following steps:

History

History is not usually included in the literature review in most of the works that you read. However, we want you to include this in your literature review since it provides a better context for a reader who is not familiar with the area of your research project. In this part, you explain the origin of the problem, sometimes even before the invention of computer systems. For instance,

> *"The first numbering system was proposed in 1st and 4th centuries by Indian mathematicians. Then, it evolved in Arabic mathematics that end up with the current decimal numbering systems."*

To my knowledge, many problems that we currently try to solve with computer algorithms and statistical methods are rooted in ancient civilizations. Anyhow, you have to dig into the history and find the first root or signs of the problem. If you do backtrack on current papers in your research project area, you can gradually find the original sources.

Hierarchy (Categories and Subcategories)

Commented [A1]: Add a sample tree structure like the one in ijcai paper

Literature Review is the formal presentation of the solutions that have been proposed to solve a problem, similar or related problems (similar ≠ related.) At the proposal step, we already explained how to explore the existing methods or similar methods to come up with a reason or motivation for the proposal. The literature review section includes all such searches in a formally *categorized* presentation.

The 1st step of the Literature Review is to create a hierarchy of solutions that have been proposed to address the problem you choose to solve. The hierarchy would have two or three levels. The hierarchy would be based on

1) the type of solutions for the problem, 2) the domain that the problem has been investigated, 2) the information sources that have been utilized, or 3) from other perspectives. What follows are some samples.

> [In terms of the type of the solution]
> *"Language modeling can be categorized into two groups: 1) n-gram language models and 2)neural language models."*
>
> *"Word embedding methods could be categorized into 1) sparse vector representation or 2) dense vector representations."*
>
> [In terms of the domain]
> *"Language modeling, based on the underlying genre of communication, would be categorized into 1) formal, 2) scholarly, and 3) informal documents."*
>
> [In term of information sources]
> *"Based on the type of information that employs, language modeling would be categorized into 1) those that use part of speeches (POS), 2) those that use syntactic information, and 3) those that use none."*

You can choose one perspective at the highest level of the hierarchy. Then, at each branch of the first level, you can continue with the second perspective. Here is an example:

> *"Based on the type of document, language modeling methods can be categorized into formal, scholarly, and informal documents. Language models for formal documents can be further divided into n-gram language models and neural language models. However, language modelling methods for informal documents are all based on neural methods."*

Usually, **up to two levels of related works are enough** for a research project unless you want to work on a survey. A survey explores all the existing methods for a problem from different angles and systematically compares them.

Summaries

The 2nd step of the Literature Review is to explain the proposed methods in each **subcategory**. Usually, time is a good hint to start with what paper. As seen in research domains, progress happens gradually, and new direction of research appear over time. So, at first, there might be only one category and only one domain. Then new category of methods emerged for a new domain. For instance:

> *"Foremost, n-gram language model was proposed and only for formal documents."*

At this point, briefly explain the best or the most well-known methods in this category. I would put at least 3 and at most 5 well-known or well-established methods:

1) their method
2) how they improve upon each other (what were the gaps of the previous one that the next one improved or filled the gaps)
3) the dataset the used

Then, you will move to the second subcategory (in our example, second subcategory of n-gram language models, i.e., n-gram language models for scholarly documents.) But before that, you have to explain why this new category emerged in the research community:

> *"N-gram language model for formal documents lacks efficiency and fall short in scholarly papers since it does not consider the authors, keywords in papers, date of publication, and the reference network. As a result, n-gram language models that incorporate them proposed."*

Commented [A2]: Add a sample table to compare works like the one in ijcai paper

Now, you continue by the similar way you did for the first subcategory, i.e., the summary of the best or well-known methods in this subcategory.

Position Your Research

The 3rd step of the Literature Review is to position your idea or your proposed solution in the hierarchy of the related words. In other words, what category and subcategory is your problem definition and proposed method. Then, you must come up with a reason why the existing methods in the subcategory fall short of solving your problem. For instance,

> *"Our proposed method is to address language models of conversations in social media. N-gram language model such as Ahmad et al. [*] and Hossein el al. [*] that proposed for formal documents perform poorly for informal documents in social media. Neural language models such as John et al. [*] and *** that could obtain state of the art performance in informal documents of social media have not been trained on conversations. Such work assume that documents happen in isolation. However, conversational documents are stream of informal documents that ***."*

This section is the most important section that strongly **reiterates** why you want to do this research.

In summary, the Literature Review has (%marking schema for this milestone):

1. (10%) History
2. (30%) Hierarchy
3. (40%) Summaries of each subcategory
4. (20%) Position the current research

Hint: Finding **a survey paper** in your research project area would make your task very easy!

Submission Guidelines

- Submission should start with the history (-∝-1960), continue with existing methods in the recent history (1960-2010) and land with state of the arts (2010-2022).
- Submission should include summaries of at least 15 existing methods: 2 in history, 3 in recent history, and 10 in the modern era.
- Submission must be written in English, in the current ACM two-column conference format in LaTeX. Overleaf templates are available from the ACM Website (use the "sigconf" proceedings template).
- Submission must be 2 pages (4 columns) in length, no more not less, including figures, tables, and references, authored by the team members. References should not be more than one column.
- Submission must be in one single zip file Literature_Review_{firstname1}_{firstname2}.zip, including:
  1. the LaTeX files
  2. the pdf file

A sample submission has been attached to this manual in Brightspace, also available online here.