Research Statement of Purpose

Abhik Jana (Department of CSE, IIT Kharagpur)

I got B.Tech in Computer Science & Engineering from Institute of Engineering & Management in 2011 and M.Tech in Computer Science & Engineering from Indian Institute of Technology (IIT) Kharagpur in 2013. Then I worked in Netapp India Private Limited in a research group (Advanced Technology Group) for 13 months. I am currently working as a PhD student in Department of Computer Science and Engineering IIT Kharagpur. My research interests lie broadly in the area of Natural language Processing, Cognitive Science.

Prospective research problems I would like to contribute:

Broad Area of Research: My PhD thesis would be based on designing an efficient way for Word Sense Induction. For this purpose we will be using the Wikipedia data to work on and we will mainly focus on Wikification. Wikification has various useful applications, for example:

- **Constructing Semantic Web**: The vision of the Semantic Web is to have semantic annotations readily available inside the webpages. The annotations produced by the Wikification process can be used to automatically enrich online documents with references to semantically related information, which will definitely improve the Web users' overall experience.
- **Educational applications :** It is important for students to have fast access to additional information relevant to the study material. The Wikification process could serve as a convenient gateway to encyclopedic information related to assignments, lecture notes, and other teaching materials, by linking important terms to the relevant pages in Wikipedia or elsewhere.
- **Natural Language Processing applications**: A number of text processing problems are likely to find new solutions in the rich text annotations. Wikipedia has already been successfully used in several natural language processing applications and we believe that the automatic Wikipedia-style annotation of documents will prove useful in a number of text processing tasks such as e.g., summarization, entailment, text categorization, knowledge acquisition etc.

State of the art of Wikification problem:

- There are generally two broad approaches to the Wikification problem:
    - **Local algorithms** which labels the terms in a document one by one using the local context of each term only; Most local algorithms use bag-of-words similarity between the context of the target term and the context of each candidate sense to identify the correct sense. But The bag-of-words model doesn't work well because:
      1) many words themselves are ambiguous and don't offer distinctive signals,
      2) it ignores other MWEs in the context or article by only treating them as individual words, leading to misunderstanding of some of the MWEs
    - **Global algorithms** which use global information of the sense configuration in the whole sentence to improve the previous local methods**.** Global information is the relation or constraints between Wikipedia concepts in the whole sentence; it usually comes from the context similarity between two Wikipedia articles. The context can either be the content of the article or the surrounding words of links which point to the article. While the link structure is an important source to extract relation among

concepts, the above approach misses out a more direct and accurate source of information, which is the co-occurrence between Wikipedia concepts in the corpus.

- Several recent attempts were made to combine both **local** and **global** information in Wikification like Wiki Machine based on an SVM model by combining local and global kernels.

All those state-of-the art approaches for Wikification are giving a decent level of precision, but those Wikification are for standard settings like for long, well described documents e.g. Wikipedia articles, some general text documents etc. So we will extend the Wikification task for new settings.

Task 1. Handling emerging senses of entities is one of the challenges in Wikification. It has been observed that the sense of entities get changed over time and sometimes a new sense of a particular entity can emerge. It is quite possible that for the new sense of a particular entity there is no Wikipedia page to refer to. We have a plan to work on this particular problem by taking care of the new senses of an entity.

Task 2. We have also planned to work on Wikification of technical document. For a motivational example, consider that I want to read the technical paper, "Data Modelling Considerations in Hadoop and Hive"(http://support.sas.com/resources/papers/data-modeling-hadoop.pdf). When I read the introduction, there are a lot of keywords that I am not aware of. For example, if I want to know about Hadoop Distributed File System (HDFS), I will Google it and one possible source would be Wikipedia.It would be very useful for beginners, however, if a link can be provided to http://en.wikipedia.org/wiki/Apache_Hadoop#HDFS.

There are challenges involved. In this example, there is no independent Wikipedia entry for HDFS. Also, the title page could have been different in Wikipedia and we might have to disambiguate for general articles. It will also be interesting to identify other domains (including scientific), where Wikification would be helpful.

Task 3. There are other aspects of our work, like semi-automatic creation/update of Wikipedia pages for computing fields. We will use dumps/edit history to understand how and when a new scientific paper results in an update in Wikipedia.

Task 4. Right now we are considering only Wikipedia Entities as knowledge in order to wikify some text documents. We can incorporate relations(Freebase or Dbpedia) , along with Entities to improve the Wikification process . We can look into metadata also to get some signals to wikify the document. There are also challenges to wikify nosy document like table+text, html etc. We can try to build a robust universal Wikification system which will take care of all these challenges.

References.

[1] *Local and Global Algorithms for Disambiguation to Wikipedia,Lev Ratinov, Dan Roth, Doug Downey, Mike Anderson. In* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1375–1384, Portland, Oregon, June 19-24, 2011. c 2011 Association for Computational Linguistics.

[2] Wikification via link co-occurrence, Cai, Z., Zhao, K., Zhu, K., & Wang, H. (2013). Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (CIKM '13), ACM, New York, pp. 1087–1096**.**

[3] Wikify!: linking documents to encyclopedic knowledge, R. Mihalcea and A. Csomai. In CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pages 233–242, New York, NY, USA, 2007. ACM.

[4] Relational inference for Wikification, Xiao Cheng and Dan Roth. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Wash., 18–21 October 2013, pages 1787–1796.