January 30, 2012

Research Statement

Respected Sir/Madam,

I have recently completed PhD in Computer Science from University of Grenoble, France.

Doctoral thesis
MathNat[1] (**Math**ematics in controlled **Nat**ural language) is a system which tries to automatically formalize the language of mathematics. The ultimate objective of this project, of which my thesis is a part, is to investigate how to check mechanically the validity of a mathematical text on a machine. This involves, among other things, translating an informal mathematical text into a formal text that can be understood by a proof assistant, theorem prover, or some similar system, to be verified or validated mechanically.

The system MathNat provides a controlled language having a look and feel of textbook mathematics. It also support miscellaneous linguistic features[2] to make it natural and expressive. A number of transformations are further applied on it to completely formalize it. An overview of this work is reported in [HR10b]; while a formal language to write mathematics is proposed in [HR10a]. The homepage dedicated to this work is given here[3] and the web interface of MathNat system is available here[4].

I was also involved in following projects. All of these resources are released as open-source, and therefore, freely available on the web.

The first project is my Master thesis titled "Urdu Morphology, Orthography and Lexicon Extraction". In this work, I reported a suite of resources including a fairly complete Urdu morphology, a lexicon and a small fragment of syntax (published in [HHR07]). A homepage is dedicated to this work which is given here[5].

The second project is based on the above work and I was involved as a second author and a senior developer. It is an elementary open-source Urdu grammar under GF resource grammar library [Ran09, RAB+10]. It is reported in [VHR10].

The third project is related to Punjabi. It is about the development of Punjabi morphology, corpus and lexicon. This morphology is written in GF [RAB+10]. Half of the corpus is boot-

[1] Homepage: `www.lama.univ-savoie.fr/~humayoun/phd/mathnat.html`

[2] Such as anaphoric pronouns and references, rephrasing of a sentence in multiple ways producing canonical forms, proper handling of distributive and collective readings, etc.

[3] `http://www.lama.univ-savoie.fr/~humayoun/phd/mathnat.html`

[4] `http://www.lama.univ-savoie.fr/~humayoun/imathnat/`

[5] `http://www.lama.univ-savoie.fr/~humayoun/UrduMorph/index.html`

strapped from Wikipedia and the lexicon is extracted semi-automatically. These resources are reported in paper [HR10c] and available here[6].

The fourth project is based on the third project. Similar to the second project, it is an elementary open-source Punjabi grammar under GF resource grammar library [Ran09]. Again, I was involved as a second author and senior developer. It is reported in paper [VHR11].

In most of these projects, I have **built corpora** from online texts[7], and **extracted lexicons semi automatically**.

The impact of these resources could also be assessed by the fact that these are (yet partially) used by Apertium which is an open-source machine translation system[8][9]. Apertium community actively participates in 'Google code-in' and 'Google summer of code' contests. Two projects based on my work (one for Urdu and one for Punjabi) are available for these contests[10] since 2010.

Besides these projects **an eight month internship at Xerox research center France**, gave me an exposure to the industrial research related to the data mining and software development field. My work at Xerox was to design:

- A data warehouse capable of storing very large amount of real time dataset associated with a fleet of Xerox high volume printing devices.
- Extraction Transformation and Load service (ETL) for this data warehouse.
- Accommodation of the implementation of various fault prediction algorithms to this service.
- GUI widgets enabling visualization of this data in form of graphs and charts.

During these projects, **I have extensively programmed in languages** such as Haskell, Java, C# and Grammatical Framework. However I am also familiar with some other languages mentioned in my CV.

In my recent professional activities, I was a member of a workshop organizing committee[11]. I also reviewed papers for a conference[12] which is held in India.

Finally, I am passionate about teaching and I would like to transmit my knowledge by teaching university classes. For instance, I can teach subjects such as *Computational Linguistics*, *Human/Natural Language Processing*, *Functional Programming Languages*, *Logic and*

[6] `http://www.lama.univ-savoie.fr/~humayoun/punjabi`

[7] With the tasks such as ripping wikipedia, news websites, blogs, etc.

[8] Apertium Machine Translation System:`www.apertium.org`, Wiki:`wiki.apertium.org/wiki/Main_Page`

[9] Apertium, Urdu and Punjabi resources: `http://wiki.apertium.org/wiki/Hindi_and_Urdu`, `http://wiki.apertium.org/wiki/Specific_resources_per_language`

[10] `http://wiki.apertium.org/wiki/Task_ideas_for_Google_Code-in`

[11] Coling Workshop on South and Southeast Asian Natural Language Processing (WSSANLP). Collocated with COLING 2010, Beijing, China. `http://www.sanlp.org/wssanlp/`

[12] International Conference on Information Systems for Indian Languages, India. `http://www.icisil2011.org`

Type Theory, *Machine Translation*, *Artificial Intelligence*, *Compiler Construction*, *Databases*, *Controlled languages for mathematics and software specifications*, etc. Furthermore, I can indeed teach the other courses of computer science as well.

I would also like to extend my doctoral work further as a long term research project. For that I can offer various MS/BS thesis projects to the students. For the moment, I have 20 months teaching experience to undergraduate college students at Computer Center, Govt. College for Women, Baghbanpura Lahore, Pakistan.

Thanks for your consideration. I look forward to hearing from you soon!
Sincerely,
–
Muhammad Humayoun

References

[HHR07] M. Humayoun, H. Hammarstrm, and A. Ranta. Urdu Morphology, Orthography and Lexicon Extraction. In Ali Farghaly and Karine Megerdoomian, editors, *Proceedings of the 2nd Workshop on Computational Approaches to Arabic Script-based Languages*, pages 59–68. Stanford, California, 2007.

[HR10a] M. Humayoun and C. Raffalli. Mathabs: A representational language for mathematics. In *Proceedings of the 8th International Conference on Frontiers of Information Technology. Pakistan.* ISBN: 978–1–4503–0342–2, ACM, 2010.

[HR10b] M. Humayoun and C. Raffalli. Mathnat - mathematical text in a controlled natural language. In Alexander Gelbukh, editor, *Special issue: Natural Language Processing and its Applications, Journal on Research in Computing Science*, volume 46. 2010. ISSN 1870-4069.

[HR10c] M. Humayoun and A. Ranta. Developing Punjabi Morphology, Corpus and Lexicon. In Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto, and Yasunari Harada, editors, *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 163–172. Tohoku University, Japan, 2010. ISBN 978–4–905166–00–9.

[RAB+10] A. Ranta, K. Angelov, B. Bringert, H. Burden, H. J. Daniels, M. Forsberg, T. Hallgren, H. Hammarström, K. Johannisson, J. Khegai, P. Ljunglöf, and P. Mäenpää. Grammatical Framework, Version 2.7. `http://www.grammaticalframework.org/`, 1998-2010.

[Ran09] Aarne Ranta. The GF Resource Grammar Library: A systematic presentation of the library from the linguistic point of view. *Linguistics in Language Technology, 2(2)*, 2009.

[VHR10] Shafqat Mumtaz Virk, Muhammad Humayoun, and Aarne Ranta. An open source urdu resource grammar. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 153–160, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[VHR11] Shafqat M. Virk, M. Humayoun, and A. Ranta. An open source punjabi resource grammar. In *Recent Advances in Natural Language Processing (RANLP 2011). to appear*, 2011.

These articles can be downloaded from my homepage: `http://www.lama.univ-savoie.fr/~humayoun`