

## Dieu-Thu Le

---

### PERSONAL INFORMATION

Address: Trento, Italy  
Date of birth: 17/06/1986  
E-mail: [dieu.thu@gmail.com](mailto:dieu.thu@gmail.com)  
Website: <http://disi.unitn.it/~dle/>

### RESEARCH INTERESTS

Natural language processing, text mining, knowledge discovery from data, topic modeling, image processing

### EDUCATION

- **PhD candidate in Computer Science (11/2010 - current)**  
University of Trento, Italy
- **European Masters Program in Language Communication Technology - double degree (9/2008 - 8/2010)**  
Grade: 110/110 cum laude
  - First year (9/2008 - 7/2009): Nancy 2 University, France.
  - Second year (10/2009 - 8/2010): Free University of Bozen-Bolzano, Italy.Degree focus: Logic and statistics for natural language modeling; corpus linguistic, linguistic resources and normalization; theory of computing; web technologies; similarity search
- **Bachelor of Information Technology (Computer Science), Honors Program (9/2004 - 8/2008)** College of Technology, Vietnam National University  
Grade: 8.37/10 (among top 10 out of more than 500 graduates)  
Degree focus: Computer Science, Information System

### ACADEMIC EXPERIENCE

**11/2010 - present: Trento University, Italy** (PhD student)

- “Action recognition in still images using language models”
- “Query classification via hidden topic analyses” and “Named Entity recognition in queries” in Galateas project<sup>1</sup>

**10/2009 - 8/2010: Free University of Bozen-Bolzano, Italy**

- Internship in Faculty of Computer Science: “Semantic Matching<sup>2</sup> for Classification Systems in Digital Libraries”
- Master thesis: “Named Entity Disambiguation in Digital Libraries”
  - Advisors: Raffaella Bernardi, Massimo Poesio, Patrick Blackburn
  - Description: disambiguating author name in digital libraries using language technologies (e.g., topic modeling, books clustering).
- Project in Similarity Search: implementing tree edit distance and windowed p-q gram distance in Java; doing experiments to compare the robustness of these algorithms.

**9/2008 - 8/2009: Nancy 2 University, France**

- Internship in LORIA<sup>3</sup> (summer 2009)

---

<sup>1</sup><http://www.galateas.eu/>

<sup>2</sup><http://semanticmatching.org/>

<sup>3</sup>Lorraine Laboratory of IT Research and its Applications, France

- Supervisor: Claire Gardent, Yannick Toussaint
- Topic: Recipe parsing in the Taaable project<sup>4</sup>
- Description: segmenting the recipes instructions, developing a syntactic parser for analyzing the sentences based on the assigned POS tagging, clustering ingredients according to the frequency of actions and properties that are syntactically related to them.
- Project in Web Technologies
  - Topic: An XML Keyword Search for English-French Expressions in the scientific domain
  - Description: developing a search engine for XML databases with English and French expressions in the academic domain, using XQuery, JavaScript and eXist framework. It also includes query pre-processing and fuzzy match (using the Levenshtein distance).
- Project in Wordnet
  - Topic: Representation of WordNet Synsets using Principal Component Analysis (PCA)
  - Description: for a given word, searching in Wordnet its different synsets, representing the similarity of these synsets by reducing the dimension using PCA and developing a graphical interface using Python.
- Project in NLP Applications
  - Topic: French Morphological Lexicon
  - Description: investigating and manipulating several French Morphological lexicons using Python NLTK.

**8/2006 - 9/2008: Department of Information Technology, College of Technology, Vietnam National University**

- B.S dissertation:
  - Topic: On the analysis of Large-scale Datasets Towards Online Contextual Advertising
  - Supervisors: Quang-Thuy Ha (Vietnam national university) and Xuan-Hieu Phan (Tohoku university, Japan)
- Research assistant of research group in Knowledge Discovery from Data
  - Topic: Developing content filter systems to support management and implementation public security - ensure policy; Information extraction models for finding entities and semantic relations in Vietnamese webpages
- Research assistant of research group in Grid Computing: implementing and running experiments in a distributed system

PUBLICATIONS

- Dieu-Thu Le, Raffaella Bernardi, “Query classification using topic models and support vector machine”, ACL (student research workshop), Republic of Korea, 2012
- Dieu-Thu Le, Raffaella Bernardi, Edwin Vald, “Query classification via Topic Models for an art image archive”, In Recent Advances in Natural Language Processing, RANLP, Bulgaria, 2011
- Raffaella Bernardi, Dieu-Thu Le, “Metadata Enrichment via Topic Models for Author Name Disambiguation”, Advanced Language Technologies for Digital Libraries, Hot Topic series, Lecture notes in Computer Science, Springer, 2011

---

<sup>4</sup><http://taaable.fr/>

- Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen, Susumu Horiguchi, Quang-Thuy Ha, “A Hidden Topic-Based Framework Towards Building Applications with Short Web Documents”, IEEE Transactions on Knowledge and Data Engineering. <http://doi.ieeecomputersociety.org/10.1109/TKDE.2010.27>, 2010
- Dieu-Thu Le, Cam-Tu Nguyen, Quang-Thuy Ha, Xuan-Hieu Phan, Susumu Horiguchi, “Matching and Ranking with Hidden Topics towards Online Contextual Advertising”, wi-iat, vol. 1, pp.888-891, 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008
- Dieu-Thu Le, Thi-Ngan Tran, Cam-Tu Nguyen, Thu-Trang Nguyen (2008). “A Vietnamese Ontology for semantic searching in medical domain”, The 11<sup>th</sup> National Conference on Information Technology of Vietnam, Hue, June 12-13, 2008

#### SELECTED AWARDS AND GRANTS

- **European Unions Erasmus Mundus Scholarship (2008-2010)**<sup>5</sup>  
A cooperation and mobility program in the field of higher education Scholarship for European MS Program in Language Communication Technologies (LCT)
- **Vietnamese Government Fellowship (2004-2008)**  
A scholarship given monthly by the Vietnamese Government to excellent undergraduate students in the Honors Program
- **Full-time Posco Fellowship (2005-2006)**  
A scholarship given to excellent Asian students in computer science, sponsored by the POSCO TJ Park Foundation, Korea
- **“Guong mat tre tieu bieu” award of Vietnam National University (2005-2006)**  
An award given to 5 most outstanding students in the university

#### TECHNICAL SKILLS

- Programming languages: Python, Java, C/C++, C#, JavaScript, Pascal, PHP, SQL
- Web Technologies: XML, XSLT, XQuery, DTD, CSS, Protege
- Certification: Qualification of JITEC’s Fundamental Information Technology Engineer Skill Standards, <http://www.jitec.ipa.go.jp/index-e.html>

#### LANGUAGES

- Vietnamese: native
- English: fluent
- Japanese, French, Italian: basic

#### EXTRA- CURRICULAR ACTIVITIES

- Visiting Sato laboratory in computer vision, Osaka university, Japan<sup>6</sup> (2012)
- Hadoop and Natural Language Processing summer school, Edinburgh<sup>7</sup> (2011)
- Participant in **IT Summit Program** held by Intebro (Internet Broadcasting station of Tokyo university of Technology<sup>8</sup>) in Chiba, Japan as one of the two specially invited students (2007)
- Participant in **East Asia Youth Leadership Forum**, Bangkok Thailand (2007)

<sup>5</sup><http://ec.europa.eu/education/external-relation-programmes/mundus.en.htm>

<sup>6</sup><http://www-sens.sys.es.osaka-u.ac.jp/>

<sup>7</sup><http://conferences.inf.ed.ac.uk/hadoop11/>

<sup>8</sup><http://www2.teu.ac.jp/intebro/>