THESIS INFORMATION

- The Doctoral Dissertation: **Open and Close Ontology-based Named Entity Disambiguation**
- Field: **Computer Science**
- Field code: **62.48.01.01**
- PhD student: **Nguyễn Thanh Hiên**
- Supervisor: **Assoc. Prof. Dr. Cao Hoàng Trụ**
- Institution: **Ho Chi Minh City University of Technology – VNUHCM**

1. Abstract of the dissertation

Named entities are those that are referred to by names such as people, organizations, or locations. Named entity disambiguation (NED) is a problem that aims at mapping entity names in a text to the right referents in a given source of knowledge. Having been emerging in recent years as a challenging problem, but significant to realization of the semantic web, as well as advanced development of natural language processing applications, named entity disambiguation has attracted much attention by researchers all over the world. This thesis proposes three methods for disambiguating named entities, and rigoruously investigates the three important factors affecting disambiguation performance, namely, employed knowledge sources, named entity representation features, and disambiguation models.

The knowledge sources exploited are close ontologies and Wikipedia. Close ontologies are built by experts following a top-down approach, with a hierarchy of concepts based on a controlled vocabulary and strict constraints. Wikipedia, considered as an open ontology, is built by volunteers following a bottom-up approach, with concepts formed by a free vocabulary and community agreements. The investigated features are entity names, identifiers of resolved entities, and words together with phrases surrounding a target name and surrounding names that are coreferential with that target name. Besides, the thesis exploits occurrence positions and lengths of names, and main alias of entities. This thesis proposes three models corresponding to the three above-mentioned methods: (i) a heuristic-based model; (ii) a statistical model; and (iii) a hybrid model, combining heuristics and statistics.

The common novelty of the proposed methods is disambiguating named entities iteratively and incrementally, including several iterative steps. Those named entities that are resolved in each iterative step will be used to disambiguate the remaining ones in the next iterative steps. Experiments are conducted to evaluate and show the advantages of the proposed methods. Besides, this thesis deals with the cases when entity names in text are partially recognized and entities referred to in text are outside an employed knowledge

source, as well as proposes new corresponding disambiguation performance measures.

2. Contributions of the dissertation

i). The thesis proposes a methodology that disambiguates named entities iteratively and incrementally.

ii). It proposes an ontology-based candidate ranking method that ranks candidate entites of a name using semantic relations of each candidate with identified entities around the name.

iii). It proposes a statistical candidate raking model that is applied to explore features extracted from a text, an ontology and Wikipedia by evaluating their combination in several ways. The model is also applied for NED based on an enriched ontology. From evaluation of experiment results, the thesis shows how important each feature is for disambiguation performance.

iv). It proposes a hybrid model, combining heuristics and statistics, for NED using Wikipedia. The proposed disambiguation process includes two phases; the first phase exploits heuristics to narrow down candidate entities and the second phase employs the above-mentioned statistical candidate raking model to rank the remaning ones.

v). It proposes disambiguation performance measures to evaluate NED methods that deal with the cases when entity names in a text are partially recognized and entities referred to in it are not in an employed knowledge base.

3. Suggestions for future work

- Extending the proposed ontology-based candidate ranking method and applying it to Wikipedia. In that case, semantic relations between entities are extracted from infoboxes of Wikipedia articles.
- Using some of the proposed heuristics to build a traning dataset. Then, supervised learning is employed to learn occurrence contexts of entity names, and the obtained model is applied for a new text.
- Exploiting diverse information from identified entities in a text to extend contexts for NED.

Supervisor	PhD student
A/Prof. Dr. Cao Hoàng Trụ	Nguyễn Thanh Hiên