Bachelor Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis author	Dominik Farhan	
Thesis title	Multilingual Entity Linking Using Dense Retrieval	
Year submitted	2024	
Study program	Computer Science	
Review author	doc. RNDr. Ondřej Bojar, Ph.D.	Reviewer
Department	Institute of Formal and Applied Linguistics	

Overall

	good	OK	poor	insufficient
Assignment difficulty	X	X		
Assignment fulfilled	X			
Total size *. . . text and code, overall workload*	X			

The goal of the thesis by Dominik Farhan is to develop a multilingual entity linking tool. The datasets to be used were suggested in the original assignment but the key dataset DaMuEL was not previously used for the task, so it deserved additional work. When assessing the upper bound performance that could be reached using DaMuEL, Dominik discovered interesting gaps in DaMuEL data.

The contributions are clear: a novel approach to multilingual EL, detailed analysis of training hyperparameters, very good discussion.

An important question addressed in the thesis only implicitly as part of the NIL entity linking problem (Section 2.1.5) is the question of false positives. Given the motivating example, the Greek hero Paris may be missing in the knowledge base while other Paris entities may be there. I see it as quite likely that a given Paris-hero mention would be accidentally but wrongly assigned to one of the existing Paris-person entries from the knowledge base, even in the golden-truth datasets. Arguably, Dominik is in no position to correct this. What I only find a little unfortunate is that the problem of false positives is "hidden under the carpet" (e.g. in the alias table method by returning k most common entities) instead of exposing the difficulty of selecting the correct one.

The resulting work is clearly beyond the expectations for a bachelor thesis; I expect it could be even accepted as a master thesis with some additions or further expanded explanations. The created EL system reaches performance levels comparable to other state-of-the art systems but used substantially less computational resources.

Thesis Text

	good	OK	poor	insufficient
Form *. . . language, typography, references*	X			
Structure *. . . context, goals, analysis, design, evaluation, level of detail*	X			
Problem analysis	X			
Developer documentation	X			
User Documentation	X			

The thesis consists of about 50 pages of text, totalling 70 pages including appendices.

The thesis structure is clean and easy to follow. The English text reads very well and contains only exceptionally rare errors (e.g. “It is ... *that* our primary interest lies.” instead of “where” on page 7 or “concern itself with the” instead of “concern the...” on page 34) or formulations not clear upfront (e.g. an example of DaMuEL data in Section 2.2.5 would really help).

I value high the survey of entity linking methods and datasets (Chapter 2, 10 pages), with just a small remark that the state-of-the-art GENRE method is described somewhat too succinctly and difficult to understand without an illustration. However this causes only little harm because Dominik’s approach is not based on GENRE.

The core of the thesis, Chapter 4, very compactly describes a number of decisions and advanced training tricks, e.g. the need for softmax scaling for the speed of learning. Many of these decisions are associated with various hyperparameter value choices and Chapter 6 provides a very useful summary of the carried out exploration.

I noticed some wrong boldfacing in Table 6.6 (English String similarity R@10 should be bold; which also conflicts with the discussion in Section 6.4.1 where only Japanese is reported as the language benefitting from string similarity).

Minor typesetting issues include the use of plain quotes (") instead of the curly ones (“”), and sometimes inconsistent title casing (e.g. “Results and discussion” should be capitalized).

Thesis Code	good	OK	poor	insufficient
Design *... architecture, algorithms, data structures, used technologies*	X			
Implementation *... naming conventions, formatting, comments, testing*	X			
Stability	X			

For an experimental type of thesis, the source code quality is not of the topmost priority. The submitted thesis goes the extra step and delivers a clean codebase with sufficient documentation (although e.g. the presumably unit tests in the directory tests are not described). Of particular interest are the efficiency considerations of Dominik Farhan. The fact that efficiency is a critical element in the planning was exposed by the large test set size and primarily the huge size of the candidate entities collection. Dominik has responded to this challenge very actively and at each step of the process, as apparent, i.a., from the discussion of LaBSE vs. LEALLA memory vs. speed in Section 6.2.

I believe that after fixing the 7% gap in DaMuEL data and hopefully indeed matching the state-of-the-art performance, the thesis deserves to be published as a research paper, for its gains in efficiency.

The stability of the code has been tested in depth via the many experiments.

Overall grade Excellent
Award level thesis No

Date

Signature