NUI Galway
OÉ Gaillimh

Insight

Doctoral Thesis

Monolingual Alignment of Word Senses and Definitions in Lexicographical Resources

Sina Ahmadi

M.Sc., Paris Descartes University, 2017
M.A., Sorbonne Nouvelle University, 2016
B.Eng., University of Kurdistan, 2014

Supervisor
Dr. John P. McCrae

External Examiner
Prof. Marie-Claude L'Homme

Internal Examiner
Dr. James McDermott

Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

in the
School of Computer Science
College of Science and Engineering
National University of Ireland Galway

Spring 2022

ABSTRACT

Dictionaries are fundamental resources for people to learn and document languages as well as for computers to process natural languages. A dictionary provides a fine-grained structure and description of the vocabulary of a language. With decades of advances in electronic lexicography, a significant amount of lexicographical resources are currently available. Such resources are the fruits of elaborate and strenuous efforts of lexicographers and oftentimes, are costly projects to initiate and maintain. Moreover, given the increasing number of lexical semantic resources thanks to community-driven initiatives such as Wiktionary, the alignment of such resources is of importance to promote interoperability and increase their exploitation more effectively. On the other hand, the significant progress in the field of computer science, artificial intelligence and the semantic web has been tremendously beneficial to various scientific fields, particularly language technology. Therefore, there is a necessity to leverage the current techniques and resources to facilitate the automatic alignment, integration and enrichment of lexicographical data.

The focus of this thesis is broadly on the alignment of lexicographical data, particularly dictionaries. In order to tackle some of the challenges in this field, two main tasks of word sense alignment and translation inference are addressed. The first task aims to find an optimal alignment given the sense definitions of a headword in two different monolingual dictionaries. This is a challenging task, especially due to differences in sense granularity, coverage and description in two resources. After describing the characteristics of various lexical semantic resources, we introduce a benchmark containing 17 datasets of 15 languages where monolingual word senses and definitions are manually annotated across different resources by experts. In the creation of the benchmark, lexicographers' knowledge is incorporated through the annotations where a semantic relation, namely exact, narrower, broader, related or none, is selected for each sense pair. This benchmark can be used for evaluation purposes of word-sense alignment systems. The performance of a few alignment techniques based on textual and non-textual semantic similarity detection and semantic relation induction is evaluated using the benchmark. Finally, we extend this work to translation inference where translation pairs are induced to generate bilingual lexicons in an unsupervised way using various approaches based on graph analysis. This task is of particular interest for the creation of lexicographical resources for less-resourced and under-represented languages and also, assists in increasing coverage of the existing resources. From a practical point of view, the techniques and methods that are developed in this thesis are implemented within a tool that can facilitate the alignment task.

CONTENTS

DECLARATION

I declare that this thesis, titled "*Monolingual Alignment of Word Senses and Definitions in Lexicographical Resources*", is composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification.

Galway, March 7, 2022

Sina Ahmadi

ACKNOWLEDGEMENTS

Writing this section has inspired me since the very beginning. Perhaps because I was wondering how my Ph.D. journey would shape this text at the end. And here is how it goes.

This thesis is the result of my past four years work as a Ph.D. researcher. Although I am sure this work is not flawless, I must say that finishing up my Ph.D. during Covid-19 gave me immense satisfaction. Starting my Ph.D. on an exceptionally sunny day in April 2018 in Galway, I had a sweet episode of my life full of new ideas in research, rewarding experiences in engineering, papers, and conferences that were typical of Ph.D. life at the time. To my surprise, I even rarely struggled with the type of challenges that some of my fellows complained about, such as making progress, carrying out experiments, catching deadlines to submit papers, or having a life while doing a Ph.D. However, the impact of the physical restrictions and psychological burden that Covid-19 imposed on everyone, including myself, were undeniable.

2020 was a particularly harsh year for everyone but even worse for international postgraduate students living in Ireland. Doing a Ph.D. during Covid-19 did not only require taking care of regular activities but also learning to cope in the new mainly virtual world while living in a shared house and dealing with the emerging mental health issues far away from the loved ones. Now that I look back, I feel that I should be gleeful that those daunting days are gone, despite the long-lasting effects.

Regardless of the problems, I enjoyed every single day of this journey and will always remember it with sweet memories. Living in Ireland was not without challenge, but at least, it taught me to be more grateful for having things that we usually take for granted: the generous sun, a more predictable weather, good food, fresh artisan bread, decent accommodation, museums, and a more affordable place to live. Nevertheless, I will miss this place for the friendly Irish people, the grasslands that make running more joyful, the biodiversity and the wildlife, particularly those robins singing at night!

This work could not have been accomplished without the help of many people. First and foremost, I would like to thank my esteemed supervisor, Dr. John M^{c}Crae, for his invaluable supervision, wise pragmatism, support, tutelage, and kindness during the course of my Ph.D. degree. My gratitude extends to the ELEXIS project for the funding opportunity to undertake my studies at the School of Computer Science at NUI Galway. My sincere thanks also go to Dr. Mathieu d'Aquin and Dr. Paul Buitelaar who played an important role as my graduate research committee members, and to the examination committee members, Dr. Marie-Claude L'Homme and Dr. James McDermott.

This journey could not be this memorable without my friends at Insight: Tobias Daudert, Joana Barros, Omnia Zayed, Adrian Ó Dubhghaill, Oksana Dereza, Sarah Carter, Niki Pavlopoulou, and Brendan Smith. Many thanks to the administrative staff at Insight and ELEXIS for their support: Hilda Fitzpatrick, Christiane Leahy-Coen, and Anna Woldrich. My Ph.D. experience would have certainly been different without the compassionate friends who helped me in many ways, especially by hosting me with open arms when I had difficulty finding accommodation: Housam Ziad and his lovely family, Aftab Alam and Dr. Theodorus Fransen. Likewise, thanks to Jalal Sajadi and Daban Q. Jaff for their true friendship over the years. Special thanks to Dr. Sanni Nimb who hosted me during my visit to the Society for Danish Language and Literature and the Centre for Language Technology in Copenhagen in 2019 which was an enriching experience in many ways. I would also like to warmly thank Dr. Mathieu Constant for hosting me in Nancy in 2021. Throughout my visit, I had a great time sharing my office with Pauline Gillet and Charlène Weyh.

The accomplishment of my Ph.D. studies is important to me in two other ways too. First, it marks the end of my formal education which has been continuously going on since I remember. I partially owe this to France which provided me with free education and unique, eye-opening, rich, and unforgettable experiences. Also, I am indebted to Dr. Kyumars S. Esmaili who encouraged me during my bachelor's to follow my passion for languages, linguistics, and computer science by studying natural language processing. Second, it ends as the fourth decade of my life begins, so do my plans to contribute to society and have a more significant role "to make the world a better place", something that I whimsically told my parents when I left them and I am sometimes contemplative if ever it will be eventual! I will stay optimistic.

This brings me to the most important people in my life. There is no word to properly express my gratitude to my family back in Kurdistan, particularly to my parents who always believed in me, endorsed me in my choices, and even sacrificed themselves to make sure that their children have a fair chance to achieve their goals, more than their own generation. I understand that pursuing my studies imposed thousands of kilometers of distance between us and years of untogetherness that could not have passed this way without their support and unconditional love. *Spas û xoşewîstî*! I am also deeply grateful to my family in Greece who have always supported me, have given new meaning to my life with their kind hearts, positive vibes and love. Άπειρα ευχαριστώ!

Finally, I would like to thank the love of my life, Ioanna, from the bottom of my heart for being my best friend and companion and for her patience and encouragement over the years. It is impossible to imagine this journey getting to an end without her unending support and love. To her, I would say: *Je t'aime*.