AMERICAN UNIVERSITY OF BEIRUT

TRANSFER LEARNING APPROACH TO DEVELOPING LARGE SCALE LEXICON FOR RESOURCE CONSTRAINED LANGUAGES

by
ALAA ISSAM MAAROUF

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Engineering
to the Department of Electrical and Computer Engineering
of the Maroun Semaan Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon
August 2021

AMERICAN UNIVERSITY OF BEIRUT

TRANSFER LEARNING APPROACH TO DEVELOPING LARGE SCALE LEXICON FOR RESOURCE CONSTRAINED LANGUAGES

by

ALAA ISSAM MAAROUF

Approved by:

Dr. Hazem Hajj, Associate Professor — Advisor

Electrical and Computer Engineering

Hazem

Dr. Imad Elhajj, Professor — Member of Committee

Electrical and Computer Engineering

Dr. Nizar Habash, Professor — Member of Committee

Computer Science, New York University Abu Dhabi

N Habash

Date of thesis defense: August 23, 2021

AMERICAN UNIVERSITY OF BEIRUT

THESIS, DISSERTATION, PROJECT RELEASE FORM

Student Name: Maarouf Alaa Issam

Last First Middle

☑ Master's Thesis ○ Master's Project ○ Doctoral Dissertation

□ I authorize the American University of Beirut to: (a) reproduce hard or electronic copies of my thesis, dissertation, or project; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

☑ I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes after: **One ___ year from the date of submission of my thesis, dissertation or project.**
Two ___ years from the date of submission of my thesis , dissertation or project.
Three ✓ years from the date of submission of my thesis , dissertation or project.

September 14, 2021

Signature Date

This form is signed when submitting the thesis, dissertation, or project to the University Libraries

Acknowledgements

I would like to thank my thesis advisor and committee for making this work happen. Special thanks to my committee, Prof. Imad Elhajj and Prof. Nizar Habash, for the helpful feedback and insightful discussions during my thesis proposal and defense. I would like to express my deep gratitude to my thesis advisor and mentor Prof. Hazem Hajj for his guidance and support on all levels. I have learned a lot from our technical discussions and feedback and from your kind and humble personality professor! I am lucky to be working under Prof. Hazem's supervision. I would also like to thank Dr. Gilbert Badaro for the useful discussions and for reviewing my report. Last but not least, I am immensely thankful to my family and friends for their continuous support during the journey making it smooth. Special appreciation to my parents for always encouraging me to do my best and being by my side, I can never pay them back.

An Abstract of the Thesis of

Alaa Issam Maarouf for Master of Engineering
Major: Electrical and Computer Engineering

Title: Transfer Learning Approach to Developing Large Scale Lexicon for Resource Constrained Languages

Lexical resources often form critical components in computational models for natural language processing (NLP). As a result, the pace of advances in NLP for resource-constrained languages, like Arabic, is slow due to limited resources compared to English large-scale resources such as English WordNet (EWN) which contains a rich set of semantics and relations between words. Despite progress to overcome this challenge, lexical resources for non-English languages remain limited in size and in accuracy of the semantics. In this thesis, we aim to overcome these limitations of size and accuracy in lexical resources by developing a method that generates a large-scale lexicon with rich semantics by transferring knowledge from a small lexical resource that has been reliably linked to EWN. Starting from a large-scale lexicon in the resource-constrained language without prior connections to EWN, the method aims at developing accurate links between the terms in the lexicon and EWN, thus creating the desired large-scale lexicon. While previous work had explored the link prediction problem through shallow links with limited accuracy, we focus on developing links based on deeper word semantics. We combine deep learning models with feature-based machine learning models that can benefit from the rich semantics within EWN. We propose a boosting three-step approach where we first apply transfer-learning by fine-tuning a BERT-based language model built for the low-resource constrained language followed by a decision tree classifier that uses the EWN semantics, and finally applying back-off prediction for terms with missing EWN semantics using Multilingual Universal Sentence Encoder (MUSE). The classifier predicts a link between two terms based on information from relations between the equivalent synsets in EWN using the depth of senses in the taxonomy, the number of edges separating the synsets, and the hypernym information within the is-a relationships between synsets. The first step in the boosting method aims to achieve high recall and the other two steps aim to improve precision. The proposed method is tested on Arabic to create a large-scale Arabic lexicon by predicting links between Standard Arabic Morphological Analyzer (SAMA) and EWN. For the small-scale lexicon with previously established reliable connections to EWN, we use Arabic WordNet (AWN). Compared to state-of-the-art ArSenL 2.0, the test results showed relative performance improvements in the accuracy of links with 4.1% F1 for nouns, 14.5% for verbs, and 19.1% for adjectives.

Contents

Acknowledgements **v**

Abstract **vi**

1 Introduction **1**

2 Background **5**
2.1 English WordNet (EWN) 5
2.2 Arabic WordNet (AWN) 6
2.3 Standard Arabic Morphological Analyzer 7

3 Literature Review **8**
3.1 Methods for Expanding WordNets 8
3.2 Methods for Developing Arabic Sentiment Lexicons 10

4 Challenges of Link Prediction Process **12**

5 Proposed Boosting Method for Link Prediction **14**
5.1 Problem Formulation 14
5.2 Boosting Approach 15
5.3 High-recall Link Prediction using Transfer Learning from Language-specific model (e.g. BERT) and fine-tuning with small scale lexicon 16
5.3.1 BERT Dataset 17
5.3.2 BERT Fine-tuning 18
5.4 High-Precision Link Prediction Method using EWN Semantics . . 18
5.4.1 Enriching Lemmas with WordNet Semantics 19
5.4.2 Feature Engineering from WordNet Semantics 21
5.4.3 Link Prediction Algorithm 21
5.5 Backoff Prediction for Missing EWN Semantics Using MUSE . . 22

6 Evaluation of the Proposed Approach with application to Arabic **24**
6.1 Overview of the Evaluation Process 24
6.2 Dataset Description 25
6.3 Comparison to Prior State of the Art 26

6.4 Analysis of False Negatives 27
6.5 Analysis of False Positives 30
6.6 Manual Evaluation of Predicted Links 32

7 Conclusion **34**

8 Future Work **35**

A Abbreviations **36**

Bibliography **37**

List of Figures

2.1 Synset Examples from EWN 6
2.2 Synset Examples from AWN 6
2.3 Lemma Examples from SAMA 7

4.1 Example of gold link not captured due to different word choice . . 12
4.2 Example of gold link not captured due to different inflections . . . 13

5.1 Example of EWN-AWN links 15
5.2 Overview of the Boosting Approach Applied to a Resource-constrained Language 16
5.3 Fine-tuning the pretrained BERT model 17
5.4 Workflow of Utilizing EWN Semantics for High-Precision Link Prediction 20
5.5 Example of WordNet Relations Approach on a SAMA-EWN pair 20
5.6 EWNxEWN matrices generated for all similarity measures 22
5.7 Workflow of Backoff Prediction for Missing EWN Semantics Using MUSE 23

6.1 Example of SAMA-EWN pairs from the gold data 26
6.2 Performance of proposed approach v/s ArSenL 2.0 on test data . 27
6.3 Performance of proposed approach using manual validation 32

List of Tables

3.1 Related Work on Arabic Sentiment Lexicons 11

6.1 Number of links and lemmas for nouns 25
6.2 Number of links and lemmas for verbs 25
6.3 Number of links and lemmas for adjectives 25
6.4 Performance Comparison to ArSenL 2.0, Pre = Precision, Rec = Recall 27
6.5 Results of the boosting approach on the test data for nouns . . . 28
6.6 Results of the boosting approach on the development data for nouns 28
6.7 Results of the boosting approach on the test data for verbs 28
6.8 Results of the boosting approach on the development data for verbs 28
6.9 Results of the boosting approach on the test data for adjectives . 29
6.10 Results of the boosting approach on the development data for adjectives 29

Chapter 1

Introduction

Lexical resources play a substantial role in natural language processing (NLP) tasks. As an example, sentiment analysis methods rely on lexicons to extract word-level scores [1]. Some efforts have already been placed in developing emotion classification models from text [2, 3, 4]. Since sentiment lexicons helped in improving the accuracy of sentiment classification models [5, 6], several researchers are working on developing emotion lexicons for different languages such as English, French, Chinese [7, 8, 9, 10, 11, 12, 13, 14, 15]. In the recent years, several researchers utilized sentiment lexicons to build feature-based machine learning models in conjunction with deep learning models for sentiment analysis [16, 17]. In addition, new studies revealed that lexical resources can help NLP systems be more explainable to humans [18, 19] allowing the analysis of the white-box model and providing explanations of predictions.

Challenges arise when dealing with resource-constrained languages like the Arabic language because they lack large-scale resources. To overcome these limitations, the past few years witnessed the construction of around 40 WordNets (WNs) which attempted to reflect semantic relations between words and help in NLP tasks. For Arabic, the Arabic WordNet (AWN) [20] is a hierarchical linguistic resource that follows the main structure of the English WordNet (EWN) [21]. Just like EWN, AWN's terms are semantically related by synonymy, antonymy, hypernymy and hyponymy. AWN also contains named entities and Arabic expressions (words and multi-words). Each entry in AWN consists of an Arabic lemma synset which represents a group of words with similar meaning. A synset is associated with a unique synset ID as well as semantic links to other synsets through relations like related-to, hypernyms, and hyponyms. For EWN, each synset is associated with an extended gloss which is a brief sentence explaining the meaning of the synset with examples of sentences using the synset terms. AWN has gained a lot of popularity and has become a common lexical resource for Arabic NLP tasks. The available AWN however suffers from limited coverage compared to other languages. For instance, AWN contains 9,698 synsets compared to 117,659 synsets in EWN. AWN's terms are linked to EWN by indexes that were manually validated by lexicographers. There were also attempts for developing Arabic emotion lexicons [12, 22, 23, 24]. Developing a large-scale resource for the Arabic language is in fact useful for several tasks like information retrieval [25], named entity recognition [26], word sense disambiguation [27], as well as creating sentiment lexicons which are useful in sentiment analysis tasks like the resource SentiWordNet (SWN) [28, 29]. To overcome the limitations of AWN, there has been several efforts to automatically create large-scale lexicons for the Arabic language [23, 24]. However, these methods were limited in terms of accuracy and semantics because they used surface level similarity measures for linking terms from Arabic to EWN. We overcome this issue by employing deeper word semantics in the link prediction process to generate a new state of the art Arabic sentiment lexicon. Unlike prior state of the art which used shallow similarity approaches of comparing the exact match between two words, we utilize semantic similarities to compare words' meanings rather than the morphology of words. Thus our approach detects links between words of the same meaning even if they have different word choice and shows enhanced accuracy of links on the test data compared to the prior state of the art. We evaluate our method for Arabic by enriching AWN with new links for terms extracted from a large-scale Arabic resource called Standard Arabic Morphological Analyzer (SAMA) [30].

SAMA covers an extended set of Arabic lemmas and is used in several Arabic NLP tasks. Each SAMA entry consists of the Arabic lemma, its part of speech (POS) tag, and a list of English terms that reflect the meaning of this lemma (English gloss). The desired model takes as an input a SAMA lemma and all EWN synsets, and outputs the links between the chosen SAMA lemma and all corresponding EWN synsets. What makes the problem challenging is that a SAMA lemma can map to more than one synset in EWN and vice versa. In addition, it is well known that similarity measures are key factors for successful link prediction methods. As a result, the proper choice of the similarity measure is critical. Moreover, there are several challenges associated with this problem including the highly imbalanced nature of the dataset which encompasses millions of no-links compared to few thousands of links. Therefore, the link prediction classifier should take into account the imbalanced data issue for accurately predicting the minority class constituting the links. In addition, the gloss terms of the large-scale lexicon in resource constrained language (LSLex-RCL) are limited in terms of coverage and do not include all potential meanings of the lemmas. On the other hand, EWN synsets represent a very specific meaning which might not be covered by the LSLex-RCL gloss terms due to their restricted coverage. The link prediction problem is a many-to-many where LSLex-RCL entries are more generic than EWN terms which adds complexity to the problem. It is worth mentioning that LSLex-RCL gloss terms and EWN synset terms may have different word choices which limits the accuracy of shallow similarity measures. For Arabic, gloss terms and EWN synset terms also include multi-word expressions which tend to be different between SAMA and EWN and are more difficult to match.

While previous work has explored shallow and surface-level similarity measures, we focus on developing links based on deeper word semantics. In this thesis, we aim to improve on state of the art methods that extend limited lexical resources with reliably established connections to resource-rich languages such as English. The idea is to transfer the learning from language models and from validated existing links to EWN in small-scale resources to predict new links between EWN and large-scale resource from the resource-poor language. We test our method for Arabic by extending AWN with links for terms from the large-scale Arabic resource SAMA [30]. We propose methods that combine deep learning and feature-based machine learning models that learn from the rich semantics within the WordNet's lexical taxonomy. We present a boosting approach which consists first of transfer learning and fine-tuning a language model pre-trained for the constrained language followed by a decision tree classifier that learns from features extracted from EWN's semantic relations. The third step involves a backoff method using Multilingual Universal Sentence Encoder (MUSE) [31] for predicting links in case of missing EWN semantic relations. For example, EWN adjectives do not have the is-a semantic relations which exist for nouns and verbs. In the first step, the objective is to achieve a high recall by fine-tuning the language model for the low-resource constrained language. In the second step, we aim to enhance precision by engineering features that take into account the depth of the synsets in the taxonomy, the shortest path separating them, and the is-a relationships between synsets. These features are then fed into a classifier which learns the correlation between the semantic features and absence/presence of links.

The proposed workflow is tested on Arabic to develop a large-scale Arabic lexicon by predicting links between SAMA and EWN. In comparison to the state-of-the-art ArSENL 2.0 [24], the test results demonstrated relative performance enhancements in the accuracy of links with 4.1% F1 score improvement for nouns, 14.5% for verbs, and 19.1% for adjectives. The main contributions of the proposed approach are as follows:

- Developing a new three-step boosting approach to automatically create large scale lexicon for resource constrained languages.

- A state of the art large scale Arabic Lexicon with rich EWN semantic relations.

- Deep Learning link prediction method utilizing transfer learning and fine-tuning a language model pre-trained for English.

- Link Prediction method based on feature-engineering from EWN's rich semantic relations.

- Link Prediction method for terms with missing EWN semantic relations using MUSE multilingual model.

The rest of the manuscript is organized as follows: section 2 covers background knowledge on lexical resources utilized in this work. In section 3, we highlight the literature background pertaining to previous work on WordNets expansion and development of Arabic sentiment lexicons. In section 4, we explain the challenges of the link prediction process with examples. In addition, section 5 presents the boosting approach for link prediction methodology with its main blocks. Section 6 demonstrates the superior performance of the approach by evaluating the results achieved and comparing them to previous work. Finally, sections 7 and 8 conclude the thesis work and suggest future directions respectively.

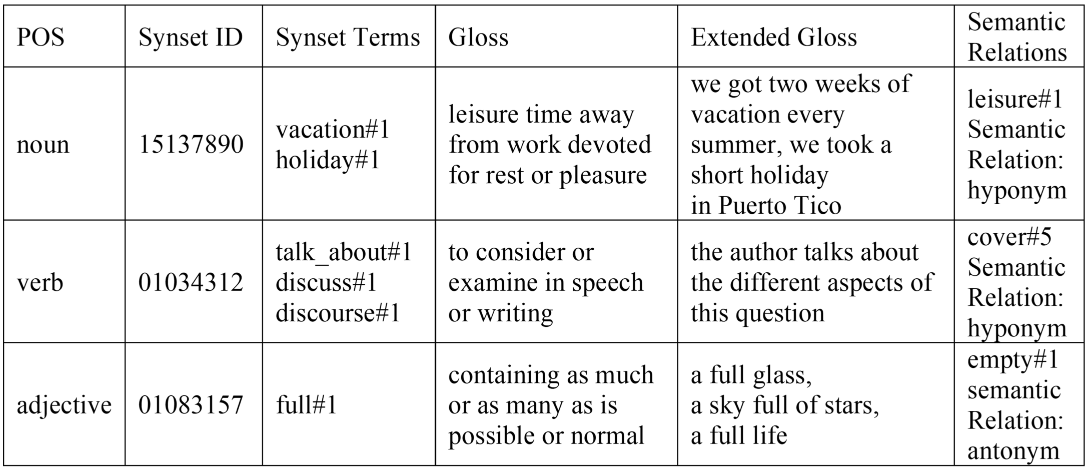
Chapter 2

Background

In this section, we cover background knowledge of the lexical resources utilized in the work.

2.1 English WordNet (EWN)

EWN [21] is a lexical resource encompassing a comprehensive set of English words which are grouped in synsets and organized hierarchically. A synset is a group of words denoting the same meaning. A unique synset ID is assigned to every synset in EWN to identify it. Every synset is associated with a gloss which provides a definition for the respective sense in addition to an extended gloss which includes examples of using the synset terms in sentences. Synsets in the WordNet are connected with relations to reflect semantics of the words like synonymy, antonymy, hypernymy and hyponymy. EWN has more than 117,000 synsets distributed among nouns, verbs, adjectives, and adverbs part of speech tags. EWN is a popular recourse for NLP tasks and has been widely used by researchers. In this work, we utilize EWN 3.0 which has 117,659 synsets entries and approximately 155,000 words. It is worth mentioning that a word can appear among multiple synsets if it has several meanings which will be indicated by a '#' symbol followed by a number denoting the corresponding meaning. Thus, EWN is a rich lexical resource and we will utilize the semantics of words drawn from this resource in our proposed approach. For instance, to compare two synsets within EWN several features can be utilized like the number of edges connecting the synsets, the hypernym/hypnonym information, and the depth of the synsets within the taxonomy. Section 5 covers more details on how these semantic relations were utilized withing our boosting approach. Table 2.1 shows examples from EWN with the aforementioned features associated with every synset.

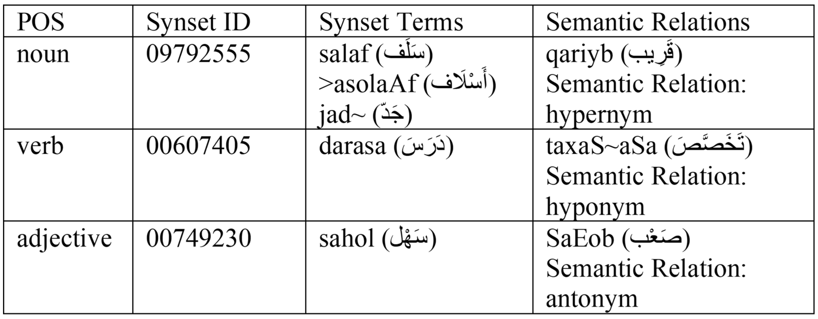


POS	Synset ID	Synset Terms	Gloss	Extended Gloss	Semantic Relations
noun	15137890	vacation#1 holiday#1	leisure time away from work devoted for rest or pleasure	we got two weeks of vacation every summer, we took a short holiday in Puerto Tico	leisure#1 Semantic Relation: hyponym
verb	01034312	talk_about#1 discuss#1 discourse#1	to consider or examine in speech or writing	the author talks about the different aspects of this question	cover#5 Semantic Relation: hyponym
adjective	01083157	full#1	containing as much or as many as is possible or normal	a full glass, a sky full of stars, a full life	empty#1 semantic Relation: antonym

Figure 2.1: Synset Examples from EWN

2.2 Arabic WordNet (AWN)

The Arabic WordNet [20] is a resource containing Arabic synsets which are semantically connected. Similar to EWN, words within a synset are synonyms sharing the same meaning. AWN also contains Names Entities and Arabic expressions (words and multi-words). In AWN, an entry consists of an Arabic lemma synset associated with a unique synset ID as well as semantic links to other synsets like related-to, hypernyms, and hyponyms. Because of the tremendous need for Arabic resources, AWN has gained a lot of popularity and became a common lexical resource for Arabic NLP tasks. The available AWN however suffers from limited coverage compared to other languages. For instance, AWN 2.0 contains 9,698 synsets compared to 117,659 synsets in EWN. AWN's terms are linked to EWN by indexes that were manually validated by lexicographers. Table 6.2 highlights examples of AWN entries for different POS tags.

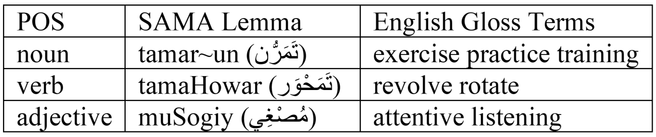


POS	Synset ID	Synset Terms	Semantic Relations
noun	09792555	salaf (سَلَف) >asolaAf (أَسْلَاف) jad~ (جَدّ)	qariyb (قَرِيب) Semantic Relation: hypernym
verb	00607405	darasa (دَرَسَ)	taxaS~aSa (تَخَصَّصَ) Semantic Relation: hyponym
adjective	00749230	sahol (سَهْل)	SaEob (صَعْب) Semantic Relation: antonym

Figure 2.2: Synset Examples from AWN

2.3 Standard Arabic Morphological Analyzer

Standard Arabic Morphological Analyzer (SAMA) [30] is a popular Arabic morphological analyzer. It encompasses a large set of Arabic lemmas that are assigned a POS and an English gloss which consists of terms indicating the meaning of the lemma. SAMA provides for an input word, all potential lemma analyses out of context. SAMA's POS tags include mainly nouns, verbs, and adjectives along with others making a total of 32 tags. Table 2.3 shows examples of SAMA entries.



POS	SAMA Lemma	English Gloss Terms
noun	tamar~un (تَمَرُّن)	exercise practice training
verb	tamaHowar (تَمَحْوَر)	revolve rotate
adjective	muSogiy (مُصْغِي)	attentive listening

Figure 2.3: Lemma Examples from SAMA

The Arabic language is morphologically rich and encompasses complex inflectional morphology which makes it challenging for several NLP tasks [32]. Moreover, Arabic words are written with diacritics which represent short vowels, nunation, and consonantal doubling. The diacritics however are not always present which leads to increased ambiguity. In addition, some Arabic letters are frequently misspelled and several variants are used to represent the same letter which adds sparsity and ambiguity to the language.

The challenges arising from Arabic's sophisticated morphology and ambiguity are usually tackled by utilizing tokenization, analysis, and disambiguation tools [33, 34]. In this thesis, we utilize the morphological analyzer SAMA [30] to identify all possible lemma readings, or analyses, for a word out of context, and thus mitigating the problem of morphological analysis.

Chapter 3

Literature Review

The objective of this literature survey is to provide a general overview about existing research related to the thesis work.

3.1 Methods for Expanding WordNets

Over the years, researchers attempted to built WordNets and expand existing ones. The authors in [35] presented an unsupervised method that utilizes word embeddings and word-sense induction to build new WordNets. The method is based on word embeddings to match words to synsets in any language having a large unannotated corpus like Wikipedia in the target language and machine translation to English. The synset information is represented using embeddings following a work on sentence embeddings [36]. The problem of cosine-similarity is erroneous, an issue which is tackled using sense clustering scheme based on Word sense Induction. The authors also create two new 600-word WordNets in French and Russian languages. In addition, the work by [37] developed a semi-automatic approach to link Hindi WordNet and the English WordNet using word embeddings. The average embedding of synset terms is computed for the source and target language to have one embedding representation for each synset per language. Afterwards, a translation vector is learned with the objective of minimizing the distance between the embedding vectors if the source and target languages. The vector can then be used to translate new synsets from the original language to the destination one and thus predict new links.

Similarly for the Arabic language, several researchers attempted to expand AWN. The authors of [38] expanded AWN using two semi-automatic techniques. The approaches utilize lexical rules that can derive related words sharing a common root. The work suggests deriving new Arabic words from existing synsets in AWN and then produce corresponding English synsets. The lexical rules used to generate the new words create regular verbal, nominal, and adjectival forms. In the first approach, the English WordNet synsets translations are obtained and then validated by lexicographers manually. The second approach suggests expanding existing synsets by deriving new Arabic forms and then obtaining English synsets for the new forms. The approach benefits from the fact that Arabic language allows for easy derivation of semantically related words from a verb root using some lexical rules. The obtained suggestions are also validated by lexicographers. Furthermore, the work in [39, 40] suggests semi-automatic techniques to expand the Arabic named entities in AWN using rich resources like Wikipedia. The goal is to extract Arabic names entities from Wikipedia and connect then as instances to existing synsets in AWN. Only names entities with English correspondents in EWN are considered. The approach utilizes Arabic resources about toponym's, countries' names, and Arabic-English lexicon for named entities. Moreover, they use Wikipedia's English and Arabic articles to create links between English and Arabic named entities. The names entities suggestions are automatically generated but are also validated manually and the words are only named entities. Moreover, the work presented in [41] combines WordNets of different languages with open licenses, data from Wiktionary and the Unicode Common Locale Data Repository to create a multilingual wordnet. The authors first built a database from existing open wordnets in more than 26 languages including Arabic. A parser was developed to parse Wiktionary data to extract headwords, parts of speech, definitions, synonyms and translations from the XML Wiktionary database dumps provided by the Wikimedia Foundation. In addition, synonyms and translations were both grouped into sense groups that correspond to definitions in the main section. These sense groups are marked by a short text gloss, which is usually an abbreviated version of one of the full definitions. The parser also generates feedback about poorly formatted data in the Wiktionary. Senses of Princeton Wordnet and Wiktionary were linked based on common translations in combination with monolingual similarity features. They use a variety of similarity sores: two of them are based on similarity in the number of lemmas calculated using the Jaccard index where the score is the ratio of the number of intersecting lemmas between Wiktionary and English WordNet over the number of the union of lemmas. They also set thresholds for filtering the results based on empirical data. They tried to improve the similarity scores by incorporating short glosses and full glosses and checking the overlap between them. Furthermore, the work in [42] follows a semi-automatic approach to expand the coverage of AWN. Named entities were extracted from YAGO and linked to their AWN synsets. The authors also linked several new nouns and verbs to extend the lexicon. The expanded AWN introduced significant improvement in the accuracy of a Q/A task.

3.2 Methods for Developing Arabic Sentiment Lexicons

Lexicons developed for the Arabic language are limited in size and coverage compared to the English language [43]. Researchers attempted to develop lexicons manually by extracting a set of words from a sentiment corpus followed by manual annotation by lexicographers [44, 45, 46]. These approaches are typically high in accuracy but suffer from limited size which is constrained by the the human effort invested in addition to being costly and time-consuming. For this reason several researchers tackled the issue by translating large-scale lexica from English to the Arabic language which was hindered by Arabic's complex morphology [47, 48]. In fact, the recent work by [49] developed multiple approaches to construct lexicons in the Arabic language and studied the impact of expanding lexicon resources on the performance of sentiment analysis. The first approach is in fact manual where linguists extract sentimental words from a specified dataset and end up with a small scale lexicon. Another approach is semi-automatic which consists of using Google Translate to translating a lexicon from English to Arabic proceeded by manual validation. In addition, the authors expand a publicly available resource resource pertaining to four different domains [50] and enrich it by generating lemmas from the lexical words using Alkhalil analyzer [51]. The lemmas generated hold multiple meanings since the words are out of context after which a manual validation is performed to filter out lemmas with opposite sentiment to the original word and neutral ones. The results showed that domain-specific lexicons improve classification accuracy when the text pertains to the same domain.

Other researchers utilized word embeddings to expand Arabic sentiment lexicons automatically. For instance, the work in [52] firstly generate a population of potential terms to be introduced to the lexicon then retrieve the top similar words using AraVec [53]. The polarity of the newly added term is calculated based on the ratio of the positive to negative sentiment of its most similar words retrieved from word embeddings and which exist in the seed lexicon. The process is recursively repeated taking the new words as the seed for additional terms. Badaro et al. [23] combined two techniques to produce the first public large-scale Arabic sentiment lexicon (ArSeL). In the first method, AWN lemmas were standardized to LDC format and mapped to English SentiWordNet (ESWN). In the second approach, the authors match gloss terms of SAMA lemmas and EWN synsets which are linked to ESWN. The resultant lexicon had the Arabic synset terms with their corresponding sentiment scores, in addition to the linked EWN synset which yielded this score. The authors then extended ArSeL to develop ArSenL 2.0 by utilizing Machine Translation (MT) tables [54] followed by surface similarity evaluation. The resultant resource is a large-scale lexicon for the Arabic language with enhanced linking accuracy compared to its predecessor ArSeL.

In summary, the previously established approaches for the Arabic language

Table 3.1: Related Work on Arabic Sentiment Lexicons

Previous Work	**Techniques**	**Automatic**	**Large Scale**	**Scalable**	**Utilizes Semantics**
BiSAL [44]	Manual annotation	X	X	X	X
[49]	Translation from English	✓	✓	✓	X
MoArLex [52]	Word embeddings	✓	✓	✓	X
ArSenL [23]	Surface matching	✓	✓	✓	X
ArSenL 2.0 [24]	MT Tables and direct matching	✓	✓	✓	X
Our Approach	Deep learning models and semantic features from EWN	✓	✓	✓	✓

rely on direct matching and surface-level similarity scores. These techniques mostly require manual validation by lexicographers. Therefore, we propose a fully automated approach to expand AWN using word semantics drawn from EWN and pre-trained language models. Table 3.1 summarizes the main work in the literature for developing Arabic Sentiment lexicons. The papers are evaluated by taking into account being manually or automatically constructed, large-scale consisting of more than 30,000 entries, scalability, ans whether word semantics were utilized. The table clearly shows that none of the existing work to automatically develop large-scale Arabic lexicons utilizes word semantics which limits the performance and accuracy, unlike our proposed approach which incorporate word semantics within the process.

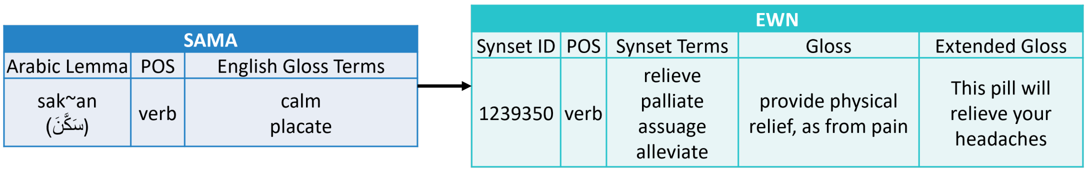
Chapter 4

Challenges of Link Prediction Process

In this work, we formulate the problem of expanding resources for a low-resource language as a link prediction problem between the gloss terms of the large-scale lexicon in the low-resource language and EWN's synset terms. Linking the two sets of words is in fact challenging as there are many constraints associated with it. In this section, we highlight the challenges pertaining to the link prediction problem and provide examples of faulty predictions from previous state of the art. This data exploration step is essential to better understand the problem and the gaps of prior work.

In what follows, we present an error analysis of the results from prior state of the art [24]. Random samples from the results of [24] were selected, and the links present in the gold data that were not predicted as absent were analyzed. The reasons for incorrect predictions is divided into two main cases each representing a different reason. The first cause for not capturing gold links was the different word choice between SAMA gloss terms and EWN synset terms. Since the work of [24] uses surface level similarity between words, it cannot match synonyms having different word choice. Figure 4.1 illustrates an example of the aforementioned problem where the SAMA lemma and EWN synset depict the same meaning yet this link wasn't captured.

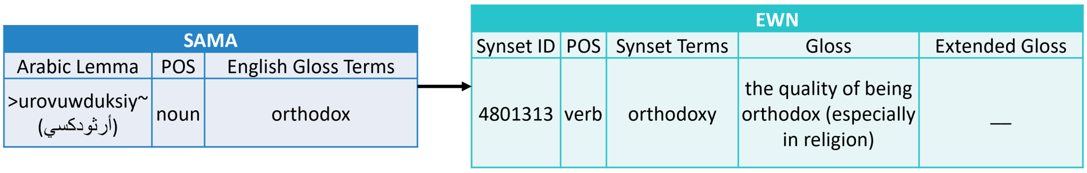
Another reason for faulty predictions is the difference in word inflections between SAMA gloss terms and EWN synset terms. Given that the work in [24]



SAMA		
Arabic Lemma	POS	English Gloss Terms
sak~an (سَكَّنَ)	verb	calm placate

EWN				
Synset ID	POS	Synset Terms	Gloss	Extended Gloss
1239350	verb	relieve palliate assuage alleviate	provide physical relief, as from pain	This pill will relieve your headaches

Figure 4.1: Example of gold link not captured due to different word choice



SAMA		
Arabic Lemma	POS	English Gloss Terms
>urovuwduksiy~ (أرثودكسي)	noun	orthodox

EWN				
Synset ID	POS	Synset Terms	Gloss	Extended Gloss
4801313	verb	orthodoxy	the quality of being orthodox (especially in religion)	—

Figure 4.2: Example of gold link not captured due to different inflections

evaluates the exact match between words, it cannot capture incompatible inflections having similar meanings. Figure 4.2 highlights an example of a SAMA-EWN link that was predicted as absent because the SAMA gloss terms and EWN synset terms do not have the same inflection.

On the other hand, our proposed approach successfully tackles the two mentioned challenges since it uses word semantics rather than surface level similarity. For instance, the examples shown in figure 4.1 and figure 4.2 were accurately captured by our boosting approach. Details of the methodology and results are explained in the following sections.

Chapter 5

Proposed Boosting Method for Link Prediction

In this chapter, we present our proposed methodology, which consists of a boosting approach encompassing several intermediate steps. The overall system utilizes deep learning as well as feature-based models to accurately link large-scale lexicons in constrained-resource languages to EWN thus creating large-scale rich semantics.

5.1 Problem Formulation

The main objective of this thesis is to develop an approach to automatically expand language resources for resource-constrained languages avoiding time-consuming manual approaches. The method is tested and evaluated on the Arabic language, and specifically on the Arabic WordNet, given its popularity and usefulness in NLP tasks [20]. In fact, the English WordNet is more than 12 times larger than the Arabic WordNet in terms of synsets coverage, which highlights the need for enriching AWN with more words and relations. To expand AWN, we enrich it with new words and predict how these words should fit in AWN's hierarchical

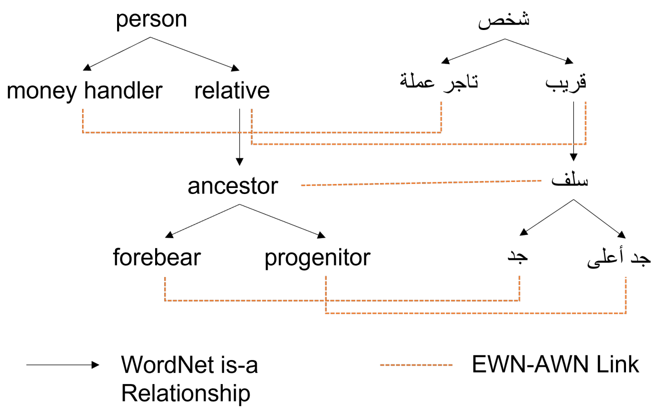


Figure 5.1: Example of EWN-AWN links

structure. For instance in Arabic, if we want to add the noun قَرِيب *qariyb* meaning *relative* to AWN, we need to predict that it is the hypernym of سَلَف *salaf* meaning *ancestor* and hyponym of شَخْص *$axoS* meaning *person* in addition to other relations. Figure 5.1 shows an example of links between EWN and AWN entries. We formulate the problem as a link prediction problem between synsets in EWN and lemmas in resource-constrained language extracted from a large lexicon without prior connection to EWN. To illustrate the method, we discuss in what follows how we achieve automated expansion for Arabic through link prediction. Since AWN is limited in size, we require an external resource with comprehensive coverage of Arabic words to be amended to AWN. In this work, we utilize the rich Arabic resource SAMA [30] to extract new Arabic words and automatically predict how these new terms will fit in AWN's hierarchical structure. SAMA encompasses 40,691 lemma/POS tag, which we refer to as Arabic lemmas. As mentioned in section 2, each lemma is associated with a set of English gloss terms which indicate the meaning of the respective lemma.

On the other hand, every synset in AWN is connected to its English counterpart in EWN through the EWN sense map files. This implies that we can infer the semantic relations of a new synset in AWN from its EWN equivalent synset. The proposed method is used to expand AWN by adding SAMA lemmas that do not already exist in AWN, for every new SAMA lemma we predict its links to EWN, thus inferring the semantic relations and the new synsets' hierarchical structure in the WordNet.

5.2 Boosting Approach

To link SAMA to EWN, we utilize the English gloss terms associated with every Arabic lemma in SAMA. From EWN, we make use of the semantic relations between synsets. We created a boosting two step approach where we first target recall by fine-tuning a language model for the resource-constrained language then precision by utilizing features from WordNet semantic relations. In the final back-off step, we utilize MUSE to link synsets with missing EWN semantics. We train and test the approach for each POS separately. Figure 5.2 highlights the high-level overview of the three-step boosting pipeline developed when applied to a resource-poor language. With application on the Arabic language, SAMA is utilized as the large-scale lexicon in the resource constrained langauge (LSLex-RCL), and the resource-constrained language (RCL) lemma is the Arabic lemma.

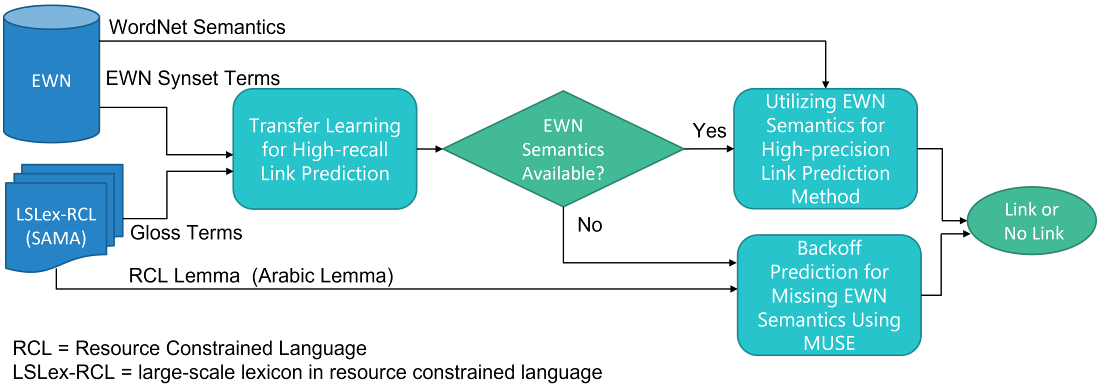

EWN
WordNet Semantics
EWN Synset Terms
LSLex-RCL (SAMA)
Gloss Terms
RCL Lemma (Arabic Lemma)
Transfer Learning for High-recall Link Prediction
EWN Semantics Available?
Yes
No
Utilizing EWN Semantics for High-precision Link Prediction Method
Backoff Prediction for Missing EWN Semantics Using MUSE
Link or No Link
RCL = Resource Constrained Language
LSLex-RCL = large-scale lexicon in resource constrained language

Figure 5.2: Overview of the Boosting Approach Applied to a Resource-constrained Language

5.3 High-recall Link Prediction using Transfer Learning from Language-specific model (e.g. BERT) and fine-tuning with small scale lexicon

In this step we utilize the pre-trained language model BERT [55] for the link prediction problem. Since the data at hand is highly imbalanced composed of no-links mostly, we first detect a significant number of no links accurately to reduce the noise coming from the numerous no-links while keeping the actual links. This step aims to achieve a good recall score with a main focus on detecting True Negatives and True Positives. However, this step induces a high number of False Positives which affects precision, which will be handled in the next step of the boosting approach. BERT [55] has been previously fine-tuned in the context of multiple language tasks among which is sentence pair classification task to detect if two input sentences are semantically equivalent. For instance, BERT proved its efficiency for the binary classification of QQP (Quora Question Pairs) by predicting whether two Quora questions are similar in terms of semantics [56]. In addition, BERT was effectively utilized for MRPC (Microsoft Research Paraphrase Corpus) to classify sentence pairs from online news websites as semantically similar or not [57]. Inspired from the two aforementioned tasks during which BERT effectively classified sentences based on semantics, we employ BERT fine-tuning for our link prediction classification task. Figure 5.3 highlights the fine-tuning process for the classification task studied. Similar to the classification for QQP and MRPC datasets, we aim to detect the similarity between SAMA gloss terms (sentence 1) and EWN synset terms (sentence 2), ultimately using semantic similarity rather than shallow one. In this step, we fine-tune BERT to predict similarity represented by link or no-link between the two sets of words. We now demonstrate how we achieved this task with having recall as the major objective.

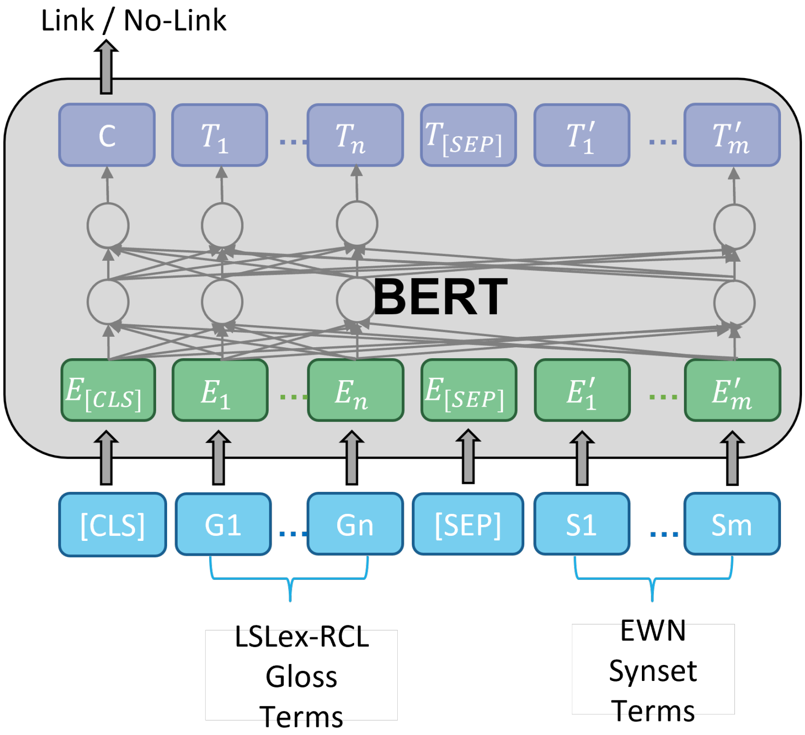


Figure 5.3: Fine-tuning the pretrained BERT model

5.3.1 BERT Dataset

As previously mentioned, the gold data was split into train, development and test data which is the same division as the work [24]. The dataset consists of comparing every SAMA lemma to the whole WordNet and predicting a binary class of link or no-link. For this step, we utilize SAMA gloss terms and EWN synset. The test data cannot be changed for a fair comparison with [24]. As for fine-tuning BERT, we use the train and development data for BERT fine-tuning and validation respectively. In fact, evaluating every SAMA lemma with all EWN synsets, yields a dataset having very few number of links compared to the number of no-links. This dataset poses a problem for BERT fine-tuning as the classification will not have enough data points from the link class to learn from compared to the no-links. We experimented with BERT fine-tuning this extremely imbalanced data, and BERT predicted no-links for all samples which is explainable for probably in a batch it had no samples or very few ones from the link class. Hence, we need to restructure the distribution of the train and development data in order for the BERT classifier to learn meaningful insights from the dataset. We constructed the BERT data to have 5% links and 95% no-links while keeping the test data intact for evaluation. In this way, we ensure that a batch will have around 5% samples from the link class to learn from. The dataset is constructed for each POS separately and we developed a structured methodology to form the train and validation data used for BERT fine-tuning. For a given POS, we take all links available links representing 5% of our target dataset. Since we would like the language model to learn that a given lemma will have links as well as significantly more no-links within the dataset, we reflect this within the resultant data. Thus, for every lemma we randomly select n no-links where n represents the 95% portion of the data.

5.3.2 BERT Fine-tuning

We utilize BERT's pre-trained language model [55] for the binary classification task. We fine-tuned the model using the dataset described above with 30 epochs and a learning rate of 3e-5. The input consists of the [CLS] token followed by SAMA gloss terms, then the [SEP] token, and finally EWN synset terms. Both SAMA glosses and EWN terms are seperated by spaces and we defined the maximum length of tokens to be 20. In addition, we added a softmax layer at the output of the model and selected the class with higher probability for each data sample. Furthermore, we selected the epoch based on the best recall score on the validation data since it is the target of this step.

5.4 High-Precision Link Prediction Method using EWN Semantics

Following the recall from the first step in the Boosting approach, we aim to improve precision in this sub-approach. We achieve enhanced precision score by retrieving all the data points that were predicted as links in the first step and filtering out the False Positives to keep the True Positives only. Since improving precision will probably decrease recall, we aim to enhance precision while keeping the recall's value reasonable and thus yielding an improved F1 score as well. As highlighted in section 2, WordNet is a rich lexical resource [21]. Every synset consists of a group of words depicting the same meanings in addition to information about hypernymns, hypononyms, and antonyms. Thus for every word with a specific meaning, one can understand its syntactic graph and how it relates to other synset in the WordNet. Since our task involves linking words based on their semantic meaning, we utilize the richness of the WordNet semantics to correlate SAMA and EWN entries.

5.4.1 Enriching Lemmas with WordNet Semantics

We developed an algorithm to link SAMA lemmas with EWN synsets. In fact, WordNet relations can be applied to synsets and not individual words. For that reason we start by mapping every SAMA entry with its corresponding synset(s). As we previously mentioned, a SAMA lemma comes with a group of English gloss terms. For every gloss term, we lookup the WordNet and retrieve the synsets where this term appears. Each synset corresponds to a specific meaning of this term. Since we don't have the context where the word appears, as we are dealing with individual words, we fetch all these synsets and associate them with the given term. The same process applies to all the gloss terms for a given SAMA lemma. Therefore, we can associate the given lemma with a group of EWN synsets corresponding to all the possible meanings of its gloss terms.

The problem is reformulated as linking a group of synsets corresponding to a SAMA entry with a synset from EWN. Since the problem is now in the WordNet space, we can utilize the rich semantic resources available within EWN. In fact, linking two synsets should take into account the semantic similarity between them including the depth of senses in the taxonomy, the number of edges separating the synsets, and the hypernym information within the is-a relationships between synsets. Thus, we utilize three different synset to synset similarity measures to reflect the aforementioned features, namely: Shortest Path, Wu-Palmer, and Leacock-Chodorow similarity. In this step, we associate with every SAMA-EWN pair a set of features built on top of the WordNet similarities. For every pair, we compute the Wu-Palmer [58], Shortest Path [59], and Leacock-Chodorow [60] similarities between the EWN synset and each of the synsets extracted from the SAMA gloss terms.

The Wu-Palmer measure computes similarity between synsets by computing the depth of the respective synsets within the WordNet taxonomies in addition to the depth corresponding to the Least Common Subsummer (LCS). The Shortest Path approach calculates the number of edges connecting two synsets within WordNet's is-a taxonomy pertaining to the shortest path. The similarity is thus inversely proportional to the number of edges constituting the shortest path.

Moreover, Leacock-Chodorow measure considers the shortest path between two synsets in addition to the maximum length of the taxonomy where the synsets are connected.

Utilizing the aforementioned similarities yields numerous features which vary in number depending on the number of SAMA gloss terms and synsets. In order to standardize these features among all data points and to organize them in a more meaningful representation, we aggregate all similarities per similarity measure into minimum, maximum, median, and average similarity. Figure 5.4 highlights the workflow for utilzing EWN semantics to achieve high precision link prediction and figure 5.5 shows an example of the approach on a SAMA-EWN pair.

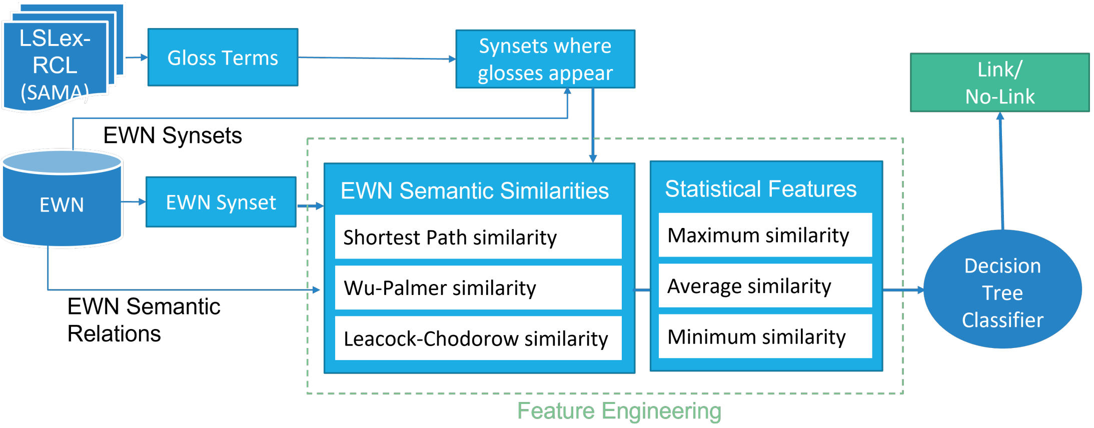


Figure 5.4: Workflow of Utilizing EWN Semantics for High-Precision Link Prediction

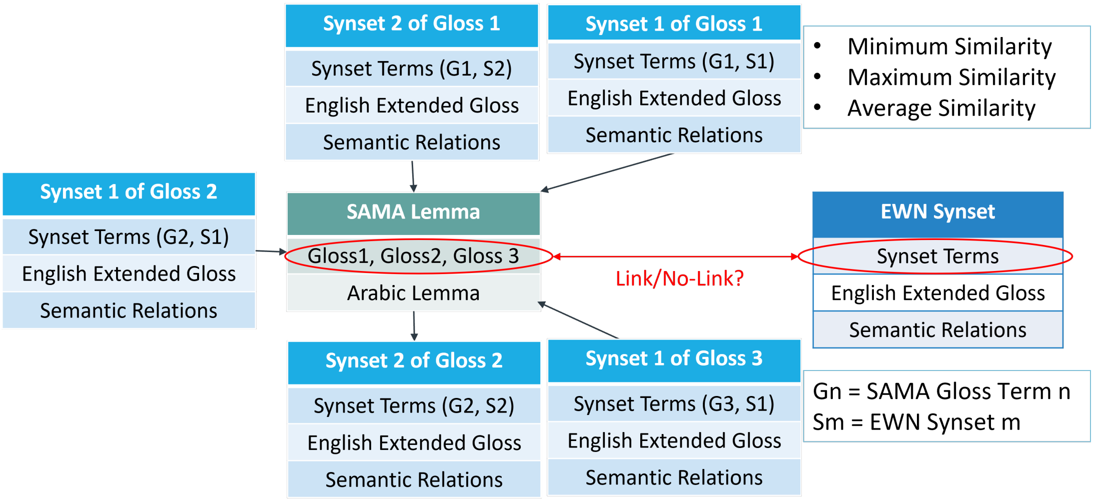


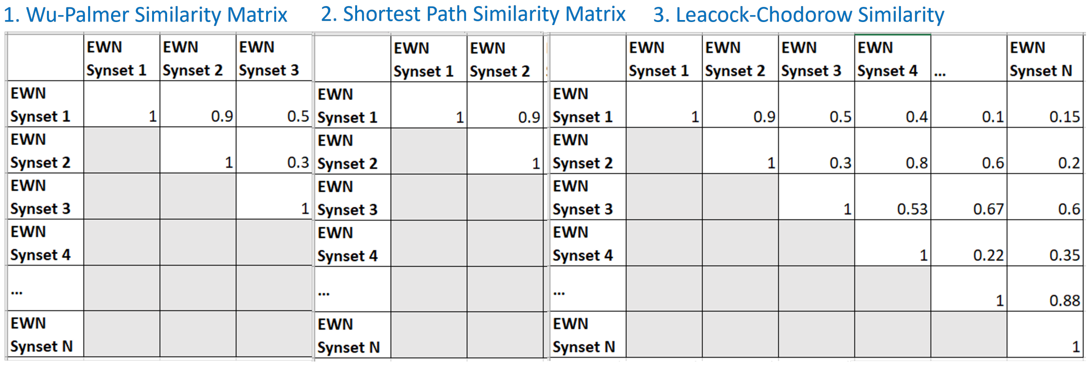
Figure 5.5: Example of WordNet Relations Approach on a SAMA-EWN pair

5.4.2 Feature Engineering from WordNet Semantics

In this section, we analyze the extracted features which consist of statistics about the WordNet similarity scores. Here we consider the case of taking one of the statistics at a time and tuning a threshold considering this feature alone to predict a link or no-link. Taking the maximum similarity alone indicates that if we identify at least one synset similar to EWN, we will consider it a link. In fact, this intuition is valid because SAMA is more generic than EWN, and if one of its meanings include EWN, we should consider it as a link. However, taking maximum similarity alone will introduce False Positives since it is an overly optimistic criteria. In addition, taking the minimum similarity signifies that even the most unrelated synset from SAMA glosses should have a minimum similarity to be considered a link. The latter is also a strict criterion if taken alone and will induce False Negatives for samples that have one of the synsets not similar to EWN but others which are. Taking average and median balances between the minimum and maximum but can be biased by outlier values or incorrect glosses. To benefit from the strengths of all these statistics and balance their effect, we take them all as numerical input features for a classifier that will predict a link or no-link. In fact, we experimented using each of the similarity measures alone verses using them all as inputs to a classifier, and the performance was boosted when having them all.

5.4.3 Link Prediction Algorithm

As highlighted in section 6.2 the dataset is large-scale with millions of records for each POS tag where the links form a very small percentage from the overall data. In fact, synsets associated with different SAMA lemmas could be repeated among lemmas. To avoid the extra computations and the repetitive ones, we generate a similarity matrix of all EWN synsets versus each other for every similarity measure. This matrix can promptly and efficiently retrieve similarities between SAMA synsets and all EWN. In addition, the matrix is in fact symmetric, so we compute only the upper half of it and then mirror it to compute the lower half. In addition, we utilize multi-processing Python libraries to run computations in parallel. The described methodology reduces the code's run-time from weeks to hours. We generate a similarity matrix for every similarity measure per POS only once, then we extract the WordNet similarity features from these matrices. Figure 5.6 shows an example of the mentioned similarity matrices. Now that the similarity matrices are available, every EWN-SAMA pair is associated with four features for minimum, maximum, average, and median for each of the three similarity measures yielding 12 numerical features. Examining the data, we notice a clear correlation between the similarity features and the output. Given that our objective is to train a classifier which can learn the thresholds ranges of the similarity scores for accurate classification prediction, we utilize the Decision



1. Wu-Palmer Similarity Matrix

	EWN Synset 1	EWN Synset 2	EWN Synset 3
EWN Synset 1	1	0.9	0.5
EWN Synset 2		1	0.3
EWN Synset 3			1
EWN Synset 4			
...			
EWN Synset N			

2. Shortest Path Similarity Matrix

	EWN Synset 1	EWN Synset 2
EWN Synset 1	1	0.9
EWN Synset 2		1
EWN Synset 3		
EWN Synset 4		
...		
EWN Synset N		

3. Leacock-Chodorow Similarity

	EWN Synset 1	EWN Synset 2	EWN Synset 3	EWN Synset 4	...	EWN Synset N
EWN Synset 1	1	0.9	0.5	0.4	0.1	0.15
EWN Synset 2		1	0.3	0.8	0.6	0.2
EWN Synset 3			1	0.53	0.67	0.6
EWN Synset 4				1	0.22	0.35
...					1	0.88
EWN Synset N						1

Figure 5.6: EWNxEWN matrices generated for all similarity measures

Tree Classifier [61] to predict links from the WordNet similarity scores. Since the objective of this step is to filter the True Positives from the links predicted by Step 1 using BERT fine-tuning, we execute the approach pertaining to this step only on the pairs predicted as links. The results clearly showed an improved precision and F1 score performance over Step 1 alone which will be further discussed in Section 6.

5.5 Backoff Prediction for Missing EWN Semantics Using MUSE

As demonstrated in section 5.4, in this work we utilize WordNet semantics to improve precision as part of the second step of the boosting approach. With regards to adjectives, more than half of the data points cannot have any of the three WordNet similarities retrieved. This is because unlike nouns and verbs, adjectives in the WordNet are not organized in a hierarchical structure [21]. For this reason, we utilize yet another model, Multilingual Universal Sentence Encoder(MUSE) [31] to enhance precision for adjectives. MUSE is a multilingual model and can be used to detect similarity between two texts from sixteen different languages. We benefit form the multi-lingual aspect of MUSE to input both Arabic and English terms.

We start by pre-processing the Arabic SAMA lemma using Arabert's preprocessor [62] followed by inputting the lemma to MUSE to compute its embedding. Afterwards, we input each of the EWN English synset terms to MUSE to get their embeddings as well. We then compute the cosine similarity between the embedding of the Arabic lemma and the embedding of each of the English sysnet terms. We can compare the embeddings between the Arabic word and the English EWN terms because MUSE maps all languages into the embeddings same space. Next, we compute the maximum similarity between the embeddings and assign this score to each data point within the dataset.

Next, we tune a threshold to decide on the cutoff bound for considering a similarity value as a link or a no-link. Similar to the WordNet approach discussed in in section 5.4, we apply this methodology on the samples predicted as links by BERT for improved precision. The reason behind tuning a threshold for the maximum similarity is that examining MUSE results we noticed that whenever the maximum similarity is lower than around 75%, MUSE predictions are not reliable which implies that the model was not trained on the specific word. However, whenever the maximum similarity is high, the predictions are reliable, and MUSE performs well which signify that the model has covered this data during pre-training. Finally, for lemmas that do not link to any synset, we compute Jaccard similarity between the corresponding lemmas and all EWN entries. The synset achieving maximum Jaccard similarity with the respective lemma is regarded as a link. If two synsets have the same maximum similarity with the SAMA lemma, they are both counted as links. Figure 5.7 highlights the workflow of the backoff prediction approach.

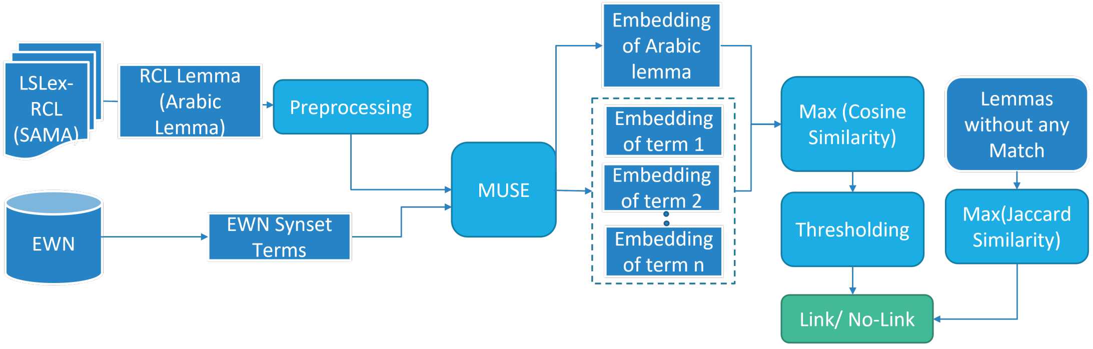


Figure 5.7: Workflow of Backoff Prediction for Missing EWN Semantics Using MUSE

Chapter 6

Evaluation of the Proposed Approach with application to Arabic

6.1 Overview of the Evaluation Process

In this chapter we evaluate our proposed boosting approach against the state-of-the-art results provided by [24]. We first evaluate our proposed approach on the development dataset to design the boosted pipeline and tune its parameters, and then run the final model on the test data for final validation. We utilize the metrics F1, precision, and recall to report performance as mentioned in 5. It is worth noting that we use these metrics because the dataset is highly imbalanced and accuracy falls short on reflecting the performance of the approach. The intuition behind improving recall is that we aim to retrieve all links available in the gold dataset. However, we also want to trust the correctness if the predicted links and avoid retrieving incorrect links which is handled by enhancing Precision. It is clear that recall and precision can be inversely proportional because retrieving more links to increase recall will introduce False Positives and deteriorate precision. Similarly, retrieving less links decreases the chance of False Positives which improves precision but decreases recall because less positive predictions will be made overall. This inverse relationship between recall and precision in fact adds complexity to the problem and clearly shows that if one method will be good on recall, it will not perform as well on precision and vice versa. This necessities the presence of a boosting approach which ensembles individual steps targeting one of the metrics each.

6.2 Dataset Description

The dataset evaluated in this work consists of SAMA lemmas and EWN synsets that are reliably linked. The data is drawn from the work of [24], where the authors constructed the dataset by taking all lemmas in common between SAMA and AWN which in turn link to EWN. The resultant dataset is referred to as the gold data. The dataset is distributed among three part-of-speech (POS) tags namely: nouns, verbs, and adjectives. The below tables show statistics pertaining to the train, test, and development datasets for the aforementioned POS tags.

Table 6.1: Number of links and lemmas for nouns

Dataset	**Number of links**	**Number of SAMA lemmas**	**Approximate number of links per lemma**
Train	7333	3562	2.06
Dev	938	445	2.11
Test	915	445	2.06
Total	9186	4452	2.06

Table 6.2: Number of links and lemmas for verbs

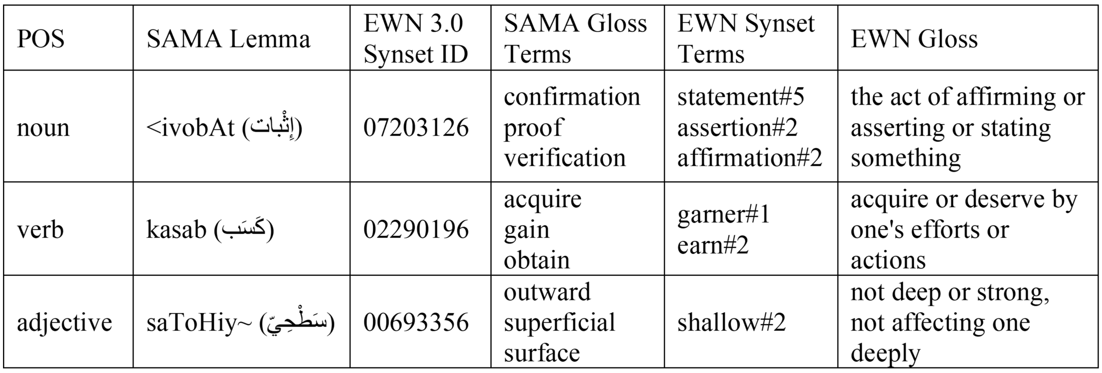
Dataset	**Number of links**	**Number of SAMA lemmas**	**Approximate number of links per lemma**
Train	4240	1724	2.46
Dev	531	215	2.47
Test	525	214	2.45
Total	5296	2153	2.46

Table 6.3: Number of links and lemmas for adjectives

Dataset	**Number of links**	**Number of SAMA lemmas**	**Approximate number of links per lemma**
Train	585	462	1.27
Dev	73	57	1.28
Test	67	57	1.18
Total	725	576	1.26

Figure 6.1 shows examples of entries from the gold set consisting of SAMA lemmas with their associated English gloss terms and the EWN synsets that they link to. Since the link prediction problem at hand is many-to-many, meaning that one SAMA lemma can link to multiple EWN synsets and vice versa, every SAMA lemma is evaluated against all EWN sysnets to detect all potential links. Comparing a lemma with all EWN yields around 2-3 links and thousands of

no-links. This makes the data hugely imbalanced and adds complexity to the problem. It is worth mentioning that SAMA lemmas are more generic than WordNet synsets, as each EWN synset represents a very specific meaning. The above tables highlight this aspect by comparing the average number of links of a SAMA entry to EWN. Every lemma links to almost 2 synsets on average which clearly demonstrates the generic nature of SAMA and adds another complexity dimension to the problem at hand.



POS	SAMA Lemma	EWN 3.0 Synset ID	SAMA Gloss Terms	EWN Synset Terms	EWN Gloss
noun	<ivobAt (إِثْبات)	07203126	confirmation proof verification	statement#5 assertion#2 affirmation#2	the act of affirming or asserting or stating something
verb	kasab (كَسَب)	02290196	acquire gain obtain	garner#1 earn#2	acquire or deserve by one's efforts or actions
adjective	saToHiy~ (سَطْحِيّ)	00693356	outward superficial surface	shallow#2	not deep or strong, not affecting one deeply

Figure 6.1: Example of SAMA-EWN pairs from the gold data

6.3 Comparison to Prior State of the Art

We argue that each step in the boosting approach aims at improving either precision or recall for an enhanced F1 score compared to [24] when all methods are ensembled. We report the results using the aforementioned metrics at every step and analyze the confusion matrix for evaluation. Results clearly show that ablating one of the mentioned steps deteriorates results, which proves that the boosting approach proposed outperforms individual methods alone. The results are reported on both development and test datasets because results on the development where utilized to design the boosting approach and the resultant ensemble model was applied on the hidden test data for final evaluation. In addition, we manually examined the results for insights and analysis.

Table 6.4 shows the results of our proposed boosting approach in comparison to ArSenL 2.0 [24] on the same test data for a fair comparison. Compared to state-of-the-art ArSenL 2.0, the test results show performance improvements in the accuracy of links with 4.1% F1 for nouns, 14.5% for verbs, and 19.1% for adjectives. We notice that our approach enhances the precision and F1 scores for all POS tags and for the average scores. We notice however a slight deterioration in recall for nouns and verbs. The improvement in precision for these POS tags however is more significant thus causing an improved F1 score for our proposed

Table 6.4: Performance Comparison to ArSenL 2.0, Pre = Precision, Rec = Recall

	Nouns			Verb			Adjective			Average		
Technique	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Our Approach	**21.81**	42.1	**28.7**	**13.8**	**36.2**	**19.9**	**38.1**	35.8	**36.9**	**17.8**	**40.0**	**24.6**
ArSenL 2.0	20.5	**42.3**	27.6	12.0	31.4	17.4	23.3	**46.3**	31	17.2	38.7	23.8

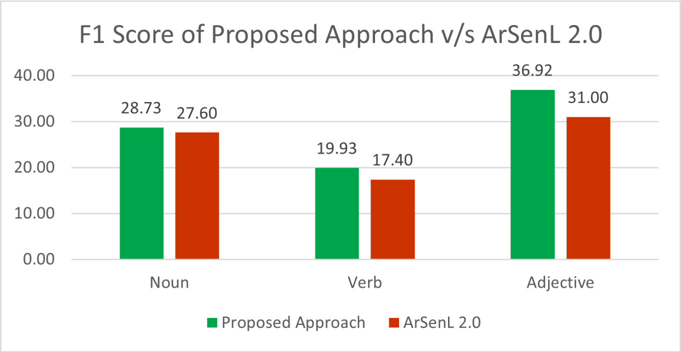


Figure 6.2: Performance of proposed approach v/s ArSenL 2.0 on test data

approach. For instance, recall for adjectives dropped by 22.6% while precision increased by 63.5%. Similarly the recall for nouns dropped by 0.5% while precision increased by 6.4%. As for verbs, both precision and recall improved by 14.5% and 14.6% respectively. Therefore, our proposed approach enhances the overall F1 score performance of the system for all POS tags.

In addition to reporting results of the overall boosting approach, we also evaluate the results of the individual sub-approaches within the boosting pipeline. In the following tables, we report the precision, recall and F1 scores for the individual steps as well as for the overall boosting approach. Results clearly show that the ensemble of the BERT fine-tuning and the WordNet relations approach combined outperform the individual methods alone for the nouns and verbs. The same applies to adjectives when BERT fine-tuning and MUSE embeddings approaches are combined. Moreover, the results demonstrate the effectiveness of step 1 and step 2 in improving recall and precision respectively thus enhancing the overall F1 score. Results are reported on the development data which was utilized to draw conclusions on the structure of the boosting approach, as well as on the unseen test data that was evaluated at the end.

6.4 Analysis of False Negatives

In this sub-section, we analyze the links that were falsely predicted as absent for nouns, verbs, and adjectives. We will now dive deeper into the reasons behind the False Negatives and False Positives. The main reason for predicting links that exist in gold as missing which are denoted by False Negatives is the limited

Table 6.5: Results of the boosting approach on the test data for nouns

Link Prediction Approach	**Nouns - TEST**				
	Method	**Objective**	**Precision**	**Recall**	**F1**
Our Approach - Step 1	**BERT**	**Recall**	0.02	75.52	0.05
Our Approach - Step 2	**WordNet**	**Precision**	15.35	52.24	23.72
Our Approach - Boosting	**Both**	**F1**	**21.81**	42.08	**28.73**
ArSenL 2.0's Approach	_	_	20.50	**42.30**	27.60

Table 6.6: Results of the boosting approach on the development data for nouns

Link Prediction Approach	**Nouns - DEV**				
	Method	**Objective**	**Precision**	**Recall**	**F1**
Our Approach - Step 1	**BERT**	**Recall**	0.03	74.52	0.065
Our Approach - Step 2	**WordNet**	**Precision**	14.96	51.28	23.16
Our Approach - Boosting	**Both**	**F1**	19.95	45.74	27.79
ArSenL 2.0's Approach	_	_	_	_	_

Table 6.7: Results of the boosting approach on the test data for verbs

Link Prediction Approach	**Verbs - TEST**				
	Method	**Objective**	**Precision**	**Recall**	**F1**
Our Approach - Step 1	**BERT**	**Recall**	0.04	89.71	0.09
Our Approach - Step 2	**WordNet**	**Precision**	11.95	38.29	18.21
Our Approach - Boosting	**Both**	**F1**	**13.75**	**36.19**	**19.93**
ArSenL 2.0's Approach	_	_	12.00	31.40	17.40

Table 6.8: Results of the boosting approach on the development data for verbs

Link Prediction Approach	**Verbs - DEV**				
	Method	**Objective**	**Precision**	**Recall**	**F1**
Our Approach - Step 1	**BERT**	**Recall**	0.04	88.51	0.08
Our Approach - Step 2	**WordNet**	**Precision**	10.58	41.24	16.83
Our Approach - Boosting	**Both**	**F1**	12.20	37.48	18.41
ArSenL 2.0's Approach	_	_	_	_	_

Table 6.9: Results of the boosting approach on the test data for adjectives

Link Prediction Approach	**Adjectives - TEST**				
	Method	**Objective**	**Precision**	**Recall**	**F1**
Our Approach - Step 1	**BERT**	**Recall**	0.03	86.57	0.05
Our Approach - Step 2	**MUSE**	**Precision**	22.05	41.79	28.87
Our Approach - Boosting	**Both**	**F1**	**38.10**	35.82	**36.92**
ArSenL 2.0's Approach	–	–	23.30	**46.30**	31.00

Table 6.10: Results of the boosting approach on the development data for adjectives

Link Prediction Approach	**Adjectives - DEV**				
	Method	**Objective**	**Precision**	**Recall**	**F1**
Our Approach - Step 1	**BERT**	**Recall**	0.01	91.78	0.02
Our Approach - Step 2	**MUSE**	**Precision**	22.35	32.11	26.36
Our Approach - Boosting	**Both**	**F1**	38.20	27.42	31.93
ArSenL 2.0's Approach	–	–	–	–	–

coverage of SAMA gloss terms. In several cases, the Arabic SAMA lemmas denote multiple meanings and thus link to multiple EWN synsets corresponding to these meanings. However, the SAMA gloss terms for these cases cover one or few of these meanings. Thus, when considering the gloss terms only from SAMA to make the link prediction, the meanings that are not covered within these terms hinder retrieving links corresponding to these missing terms.

For instance, the noun lemma تَوْدِيع *tawodiyE* with with the English gloss terms 'departure', 'farewell' is linked in the gold set to the EWN synset 372448, with synset terms 'deposit', 'deposition' and extended gloss 'the act of putting something somewhere'. This link was not retrieved by the boosting approach because the gloss terms of the Arabic lemma *tawodiyE* do not include all the possible meanings of the lemma. In other words, the link to the EWN synset 372448 corresponding to 'deposit' was not detected because the SAMA gloss terms do not cover the corresponding meaning. However, the lemma *tawodiyE* was correctly linked to the EWN synset 53097 with synset terms ‘farewell’, ‘leave’, ‘leave-taking’, ‘parting’, and extended gloss consisting of ‘departing politely’; ‘he disliked long farewells’; ‘he took his leave’; ‘parting is such sweet sorrow’. This link was captured by the proposed approach because the ‘leave’ meaning is covered within the SAMA gloss terms although not all the terms match.

As a verb example, the SAMA lemma سَكَّن *sak∼an* having the English gloss terms ‘calm’,‘placate’ is linked in the gold dataset to the EWN lemma 415828 with synset term ‘resettle’ and an extended gloss of : ‘settle in a new place’; ‘The immigrants had to resettle’. This link wasn’t detected by the proposed approach because the SAMA gloss terms cover only one meaning out of the possible meanings for the lemma *sak∼an.* SAMA gloss terms for this lemma should also include the resettling meaning of the Arabic word. On the other hand, this methodology was able to detect links from the gold set whose meaning is covered within the SAMA gloss terms, even for the cases where EWN and SAMA have different word choice for the same meaning. For example, the approach detected the link between the Arabic lemma *sak∼an* and the EWN synset 1239350 having the synset terms ‘relieve’, ‘palliate’, ‘assuage’, ‘alleviate’ and extended gloss ‘provide physical relief, as from pain’; ‘This pill will relieve your headaches’. This example clearly shows that the limited coverage of SAMA gloss terms hinders detecting links from the gold set.

Similarly for adjectives, the lemma جُزْئِيّ *juzo}iy∼* with the English gloss terms ‘in part’, ‘partial’, ‘partially’, ‘petty’ is linked in the gold set to the EWN synset 2900700 having the synset term ‘molecular’ and the extended gloss ‘relating to or produced by or consisting of molecules’; ‘molecular structure’; ‘molecular oxygen’; ‘molecular weight is the sum of all the atoms in a molecule’. Again, this is because the English gloss terms of the lemma *juzo}iy∼* do not have a comprehensive coverage of all the possible meanings of the lemma which hindered the detection of this link.

6.5 Analysis of False Positives

In this section, we explain with examples the reasons for predicting links as present while they are absent. Examining closely the results of the boosted link prediction approach, we notice that most of the links that where falsely predicted as existing in fact make sense. These links however are not present in the gold dataset because of limited coverage of links in the gold dataset. It is worth mentioning that re-evaluating the gold dataset to add all missing links will significantly improve the precision performance of the proposed approach because most of the False Positives will count as True Positives instead.

For instance, the noun lemma إِثْبات *<ivobAt* with English gloss terms ‘confirmation’, ‘proof’, ‘verification’ was predicted to link to the EWN synset 7179943 with synset terms ‘ratification’, ‘confirmation’ and extended gloss consisting of ‘making something valid by formally ratifying or confirming it’; ‘the ratification of the treaty’ ; ‘confirmation of the appointment’. Although this link does not exist among the gold data links, it actually makes sense in terms of meaning and was considered as a False Positive. The *<ivobAt* Arabic lemma however is predicted to link to the synset 5824739 having the synset terms ‘proof’, ‘cogent evidence’ and the extended gloss ‘any factual evidence that helps to establish the truth of something’; ‘if you have any proof for what you say, now is the time to produce it’. This link is considered a True Positive because it also exists in gold unlike the first case.

Similarly, the verb Arabic lemma غَسَل *gasal* with English gloss terms ‘clean’, ‘wash’ is predicted by the proposed approach to link to the EWN synset 557686 with synset terms ‘wash out’, ‘wash off’, ‘wash away’, ‘wash’ and extended gloss consisting of ‘remove by the application of water or other liquid and soap or some other cleaning agent’; ‘he washed the dirt from his coat’; ‘The nurse washed away the blood’; ‘Can you wash away the spots on the windows?’; ‘he managed to wash out the stains’. This link does not exist in the gold set although it actually makes sense in terms of meaning. This highlights the lack of complete coverage of links within the gold dataset. Thus, although this sample is tagged as a False Positive, it should be counted towards True Positives instead which significantly boosts precision if all similar cases are also counted as True Positives instead. In fact the lemma *gasal* is also predicted to link to the EWN synset 177714 with synset terms ‘pick’, ‘clean’ and with the extended gloss ‘remove unwanted substances from, such as feathers or pits’; ‘Clean the turkey’. Unlike the previous case, the link actually exists in the gold and was thus counted towards True Positives which should have been the case for the first case as well.

The adjective Arabic lemma وُدِّيّ *wud∼iy∼* with English gloss terms ‘amicable’, ‘cordial’, ‘friendly’ was predicted to link to the EWN synset 1077995 having the synset term ‘friendly’ and the extended gloss ‘easy to understand or use’;

‘user-friendly computers’; ‘a consumer-friendly policy’; ‘a reader-friendly novel’. This link does not appear among the gold links but in fact it makes sense as both the Arabic lemma and the EWN synset refer to the meaning ‘friendly’. This link was regarded as a False Positive although in fact it is a missing True Positive from the gold set.

6.6 Manual Evaluation of Predicted Links

The proposed boosting approach was utilized to predict links for the unseen SAMA lemmas that are not present in the gold data. Hence, a large-scale lexicon was generated with Arabic lemmas linked to EWN. Error analysis was conducted to validate the accuracy of the predicted links by randomly selecting 400 pairs of SAMA lemmas and EWN synsets consisting of 133 adjective, 133 verb, and 134 noun entries. Figure 6.3 highlights the accuracy of the predicted links depicted by the proportion of the correctly predicted links relative to the total number of links predicted per POS tag. The proposed boosting approach achieved an accuracy corresponding to the accurately predicted links relative to all the links predicted of 61.65%, 80.45%, and 81.34% for adjectives, verbs, and nouns respectively.

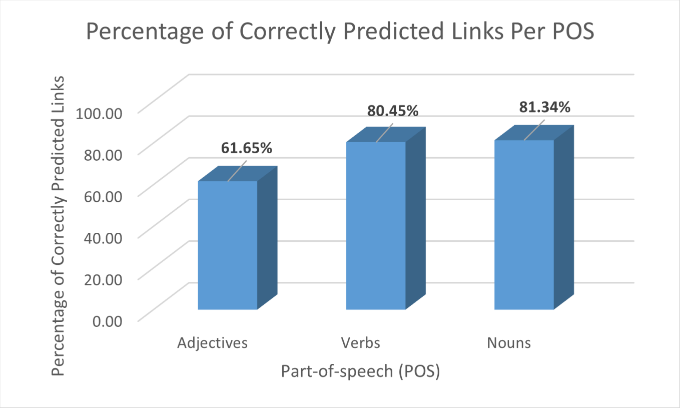


Figure 6.3: Performance of proposed approach using manual validation

Analyzing the False Positives, we notice that most of the links that were predicted as present while in fact they should be absent are actually homonyms. In other words, SAMA gloss terms and EWN synset terms may be exactly the same, but with different meanings. A potential solution is to incorporate information from EWN extended gloss and the Arabic lemma to capture the exact semantics of words having multiple meanings. It is worth mentioning that results for adjectives were the least accurate among the three POS tags because in some cases corresponding to the adjective links, antonyms were predicted as links which was not the case for verbs and nouns. For adjectives, we utilized MUSE embeddings which couldn’t always differentiate between synonyms and antonyms since they appear in the same context. However, for nouns and verbs, WordNet relations mitigates the issue of matching antonyms by relying on word semantics. Hence, the WordNet semantics approach is more reliable than the MUSE embeddings one. A potential solution is amending the adjectives within the WordNet with hierarchical relations similar to nouns and verbs followed by applying the WordNet relations approach to adjectives similar to verbs and nouns.

Chapter 7

Conclusion

The thesis provided a new transfer learning method to expand lexical resources for resource-constrained languages by combining several machine learning technologies including: transfer learning, deep learning, and feature-based machine learning. Each step in the process aimed at efficiently making use of the strength of the respective method. The first step used deep learning and transfer learning to build on the strength of universal language models. The second step employed the specific strength of the EWN resources by extracting rich semantic features and applying a shallow classifier. The backoff method utilizing MUSE aimed at closing the gap in missing EWN semantics. The work resulted in a new state-of-the-art Arabic lexicon.

Since the data at hand is highly imbalanced, composed of no-links mostly, we first detect True Negatives to reduce the noise coming from the numerous no-links while keeping the actual links. In the second step, we developed an algorithm to link terms in the large-scale lexicon in the resource constrained language with EWN synset terms by employing WordNet relations which take into consideration word semantics drawn from EWN's rich taxonomy. The features extracted from the WordNet relations are then fed to a tree-based classifier to make the final link prediction. This step improves precision by filtering out False Positives which are induced from the first step.

Compared to previous state of the art the proposed approach resulted in improved F1 score performance for noun, verb, and adjective POS tags by 4.1%, 14.5%, and 19.1% respectively. Unlike ArSenL 2.0, our approach detects links between sets of words having the same meaning even if they have different word choice. Analyzing the results of the presented methodology, we noticed that the main reason for False Positives is missing links from the gold data while the main reason for False Negatives is limited coverage of SAMA gloss terms.

Chapter 8

Future Work

For the future work, we intend to enhance the coverage of links in the gold data through manual validation. This will help in improving the accuracy of the developed models as well as the resultant F1 score since several predictions that are regraded as False Positives will count towards True Positives instead. In addition, we aim to explore different language models within the first step of the boosting approach for enhanced results. For instance we can experiment with recent language models like GPT2 [63], XLNet [64], and RoBERTa [65].

In addition, we will apply and test the proposed method to other resource-constrained languages like Hindi, Turkish, and Swahili which suffer from poor resources for NLP tasks. Since there are WordNets available for most of the languages and are connected to EWN, we can use WordNets as the seed resource for the resource-constrained language. We still need to identify a large-scale lexicon in the resource constrained language of interest, like SAMA for Arabic, to provide the new terms that will be introduced. The remaining components o the methodology can be applied seamlessly given the generalization nature of our approach. After executing the proposed boosting pipeline on other languages, we will analyze results and draw insights about the correlation between the quality of lexical resources employed in the approach, like AWN and SAMA for Arabic, on one hand and the resulting performance on the other hand.

Moreover, we intend to test the developed lexicon in the frame of various NLP tasks like sentiment analysis and evaluate the impact of utilizing the lexicon on the results. This is especially helpful when the train data is small-scale which makes it difficult to utilize language models. Since the method is automated, it can be promptly executed and tested on other languages. The developed lexicons will serve as new state of the art resources for resource-constrained languages which can be incorporated with a wide range of NLP tasks for the respective languages.

Appendix A

Abbreviations

AI	Artificial Intelligence
AWN	Arabic WordNet
BERT	Bidirectional Encoder Representations from Transformers
DL	Deep Learning
ESWN	English SentiWordNet
EWN	English WordNet
LM	Language Model
LCS	Least Common Subsummer
LSLex-RCL	Large-Scale Lexicon in Resource-Constrained Language
ML	Machine Learning
MT	Machine Translation
MUSE	Multilingual Universal Sentence Encoder
NLP	Natural Language Processing
POS	Part-of-speech
Q/A	Question Answering
RCL	Resource-Constrained Language
SAMA	LDC Standard Arabic Morphological Analyzer
SSLex-RCL	Small-Scale Lexicon in Resource-Constrained Language

Bibliography

[1] G. Badaro, O. El Jundi, A. Khaddaj, A. Maarouf, R. Kain, H. Hajj, and W. El-Hajj, “Ema at semeval-2018 task 1: Emotion mining for arabic,” in *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 236–244, 2018.

[2] S. Shaheen, W. El-Hajj, H. Hajj, and S. Elbassuoni, “Emotion recognition from text based on automatically generated rules,” in *2014 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 383–392, IEEE, 2014.

[3] A. Houjeij, L. Hamieh, N. Mehdi, and H. Hajj, “A novel approach for emotion classification based on fusion of text and speech,” in *Telecommunications (ICT), 2012 19th International Conference on*, pp. 1–6, IEEE, 2012.

[4] M. Abdul-Mageed and L. Ungar, “Emonet: Fine-grained emotion detection with gated recurrent neural networks,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 718–728, 2017.

[5] B. Liu and L. Zhang, “A survey of opinion mining and sentiment analysis,” in *Mining text data*, pp. 415–463, Springer, 2012.

[6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-based methods for sentiment analysis,” *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[7] S. M. Mohammad, “Word affect intensities,” *arXiv preprint arXiv:1704.08798*, 2017.

[8] A. Bandhakavi, N. Wiratunga, S. Massie, and D. Padmanabhan, “Lexicon generation for emotion detection from text,” *IEEE intelligent systems*, vol. 32, no. 1, pp. 102–108, 2017.

[9] C. Yang, K. H.-Y. Lin, and H.-H. Chen, “Building emotion lexicon from weblog corpora,” in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 133–136, Association for Computational Linguistics, 2007.

[10] D. Das, S. Poria, and S. Bandyopadhyay, “A classifier based approach to emotion lexicon construction,” in *International Conference on Application of Natural Language to Information Systems*, pp. 320–326, Springer, 2012.

[11] S. Poria, A. Gelbukh, D. Das, and S. Bandyopadhyay, “Fuzzy clustering for semi-supervised learning–case study: construction of an emotion lexicon,” in *Mexican International Conference on Artificial Intelligence*, pp. 73–86, Springer, 2012.

[12] S. M. Mohammad and P. D. Turney, “Crowdsourcing a word–emotion association lexicon,” *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[13] A. Abdaoui, J. Azé, S. Bringay, and P. Poncelet, “Feel: a french expanded emotion lexicon,” *Language Resources and Evaluation*, vol. 51, no. 3, pp. 833–855, 2017.

[14] J. Staiano and M. Guerini, “Depechemood: a lexicon for emotion analysis from crowd-annotated news,” *arXiv preprint arXiv:1405.1605*, 2014.

[15] G. Badaro, H. Jundi, H. Hajj, and W. El-Hajj, “Emowordnet: Automatic expansion of emotion lexicon using english wordnet,” in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 86–93, 2018.

[16] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, “Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning,” *IEEE Access*, vol. 8, pp. 23522–23530, 2020.

[17] Y. Chen, J. Xiong, W. Xu, and J. Zuo, “A novel online incremental and decremental learning algorithm based on variable support vector machine,” *Cluster Computing*, vol. 22, no. 3, pp. 7435–7445, 2019.

[18] S. M. Carta, S. Consoli, L. Piras, A. S. Podda, and D. R. Recupero, “Explainable machine learning exploiting news and domain-specific lexicon for stock market forecasting,” *IEEE Access*, vol. 9, pp. 30193–30205, 2021.

[19] V. S. Silva, A. Freitas, and S. Handschuh, “Xte: Explainable text entailment,” *arXiv preprint arXiv:2009.12431*, 2020.

[20] S. Elkateb, W. Black, P. Vossen, D. Farwell, H. Rodríguez, A. Pease, and M. Alkhalifa, “Arabic wordnet and the challenges of arabic,” in *Proceedings of Arabic NLP/MT Conference, London, UK*, Citeseer, 2006.

[21] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, “Introduction to wordnet: An on-line lexical database,” *International journal of lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[22] A. F. El Gohary, T. I. Sultan, M. A. Hana, and M. M. El Dosoky, "A computational approach for analyzing and detecting emotions in arabic text," *International Journal of Engineering Research and Applications (IJERA)*, vol. 3, pp. 100–107, 2013.

[23] G. Badaro, R. Baly, H. Hajj, N. Habash, and W. El-Hajj, "A large scale arabic sentiment lexicon for arabic opinion mining," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pp. 165–173, 2014.

[24] G. Badaro, H. Hajj, and N. Habash, "A link prediction approach for accurately mapping a large-scale arabic lexical resource to english wordnet," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 19, no. 6, pp. 1–38, 2020.

[25] R. Mandala, T. Tokunaga, and H. Tanaka, "The use of wordnet in information retrieval," in *Usage of WordNet in Natural Language Processing Systems*, 1998.

[26] M. Negri and B. Magnini, "Using wordnet predicates for multilingual named entity recognition," in *Proceedings of The Second Global Wordnet Conference*, pp. 169–174, 2004.

[27] S. Banerjee and T. Pedersen, "An adapted lesk algorithm for word sense disambiguation using wordnet," in *International conference on intelligent text processing and computational linguistics*, pp. 136–145, Springer, 2002.

[28] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.," in *Lrec*, vol. 10, pp. 2200–2204, 2010.

[29] A. Esuli and F. Sebastiani, "Sentiwordnet: a high-coverage lexical resource for opinion mining," *Evaluation*, vol. 17, no. 1, p. 26, 2007.

[30] M. Maamouri, D. Graff, B. Bouziri, S. Krouna, and S. Kulick, "Ldc standard arabic morphological analyzer (sama) v. 3.1," *LDC Catalog No. LDC2010L01. ISBN*, pp. 1–58563, 2010.

[31] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, *et al.*, "Multilingual universal sentence encoder for semantic retrieval," *arXiv preprint arXiv:1907.04307*, 2019.

[32] W. Salloum and N. Habash, "Adam: Analyzer for dialectal arabic morphology," *Journal of King Saud University-Computer and Information Sciences*, vol. 26, no. 4, pp. 372–378, 2014.

[33] N. Habash and O. Rambow, "Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop," in *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pp. 573–580, 2005.

[34] M. Diab, K. Hacioglu, and D. Jurafsky, "Automated methods for processing arabic text: from tokenization to base phrase chunking," *Arabic Computational Morphology: Knowledge-based and Empirical Methods. Kluwer/Springer*, p. 66, 2007.

[35] M. Khodak, A. Risteski, C. Fellbaum, and S. Arora, "Automated wordnet construction using word embeddings," in *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pp. 12–23, 2017.

[36] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," 2016.

[37] K. Patel, D. Kanojia, and P. Bhattacharyya, "Semi-automatic wordnet linking using word embeddings," in *Proceedings of the 9th Global Wordnet Conference*, pp. 266–271, 2018.

[38] H. Rodríguez, D. Farwell, J. Ferreres, M. Bertran, M. Alkhalifa, and M. A. Martí, "Arabic wordnet: Semi-automatic extensions using bayesian inference.," in *LREC*, 2008.

[39] M. Alkhalifa and H. Rodríguez, "Automatically extending named entities coverage of arabic wordnet using wikipedia," *International Journal on Information and Communication Technologies*, vol. 3, no. 3, pp. 20–36, 2010.

[40] M. Alkhalifa and H. Rodrguez, "Automatically extending ne coverage of arabic wordnet using wikipedia," in *Proc. Of the 3rd International Conference on Arabic Language Processing CITALA2009, Rabat, Morocco*, 2009.

[41] F. Bond and R. Foster, "Linking and extending an open multilingual wordnet," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 1352–1362, 2013.

[42] L. Abouenour, K. Bouzoubaa, and P. Rosso, "On the evaluation and improvement of arabic wordnet coverage and usability," *Language resources and evaluation*, vol. 47, no. 3, pp. 891–917, 2013.

[43] O. Oueslati, E. Cambria, M. B. HajHmida, and H. Ounelli, "A review of sentiment analysis research in arabic language," *Future Generation Computer Systems*, vol. 112, pp. 408–430, 2020.

[44] K. Al-Rowaily, M. Abulaish, N. A.-H. Haldar, and M. Al-Rubaian, “Bisal–a bilingual sentiment analysis lexicon to analyze dark web forums for cyber security,” *Digital Investigation*, vol. 14, pp. 53–62, 2015.

[45] M. Abdul-Mageed, M. Diab, and M. Korayem, “Subjectivity and sentiment analysis of modern standard arabic,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 587–591, 2011.

[46] M. Abdul-Mageed and M. Diab, “Toward building a large-scale arabic sentiment lexicon,” in *Proceedings of the 6th international global WordNet conference*, pp. 18–22, 2012.

[47] A. Mourad and K. Darwish, “Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs,” in *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pp. 55–64, 2013.

[48] M. Abdul-Mageed, M. Diab, and S. Kübler, “Samar: Subjectivity and sentiment analysis for arabic social media,” *Computer Speech & Language*, vol. 28, no. 1, pp. 20–37, 2014.

[49] I. Touahri and A. Mazroui, “Deep analysis of an arabic sentiment classification system based on lexical resource expansion and custom approaches building,” *International Journal of Speech Technology*, vol. 24, no. 1, pp. 109–126, 2021.

[50] H. ElSahar and S. R. El-Beltagy, “Building large arabic multi-domain resources for sentiment analysis,” in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 23–34, Springer, 2015.

[51] M. Boudchiche, A. Mazroui, M. O. A. O. Bebah, A. Lakhouaja, and A. Boudlal, “Alkhalil morpho sys 2: A robust arabic morpho-syntactic analyzer,” *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 2, pp. 141–146, 2017.

[52] M. Youssef and S. R. El-Beltagy, “Moarlex: An arabic sentiment lexicon built through automatic lexicon expansion,” *Procedia computer science*, vol. 142, pp. 94–103, 2018.

[53] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, “Aravec: A set of arabic word embedding models for use in arabic nlp,” *Procedia Computer Science*, vol. 117, pp. 256–265, 2017.

[54] A. El Kholy and N. Habash, “Orthographic and morphological processing for english–arabic statistical machine translation,” *Machine Translation*, vol. 26, no. 1, pp. 25–45, 2012.

[55] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[56] Z. Chen, H. Zhang, X. Zhang, and L. Zhao, "Quora question pairs," *URL https://www. kaggle. com/c/quora-question-pairs*, 2018.

[57] W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.

[58] Z. Wu and M. Palmer, "Verb semantics and lexical selection," *arXiv preprint cmp-lg/9406033*, 1994.

[59] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Computational linguistics*, vol. 32, no. 1, pp. 13–47, 2006.

[60] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," *WordNet: An electronic lexical database*, vol. 49, no. 2, pp. 265–283, 1998.

[61] N. Horning, "Introduction to decision trees and random forests," *Am. Mus. Nat. Hist*, vol. 2, pp. 1–27, 2013.

[62] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," *arXiv preprint arXiv:2003.00104*, 2020.

[63] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[64] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[65] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.