A Method for Proper Noun Extraction in Kurdish

Hossein Hassani

University of Kurdistan Hewlêr, Erbi, Kurdistan Region, Iraq; and
Sarajevo School of Science and Technology, Sarajevo, Bosnia and Herzegovina
hosseinh@ukh.edu.krd, hossein.hassani@stu.ssst.edu.ba

Abstract

This paper suggests a method for proper noun identification in Kurdish texts. Kurdish proper nouns are not capitalized and they also assume other part-of-speech roles, which leads to a broad ambiguity that should be addressed in Kurdish proper noun recognition applications. Kurdish is also among less-resourced languages. We developed an application based on an architecture which includes a number of name lists, a set of rules, and a set of processes that recognizes Kurdish person names. This can help the study of Information Retrieval (IR) in Kurdish to advance and can also be used in Kurdish machine translation. We conducted several experiments which showed that the precision of the method is more than 95%, the recall is between 40% to 80%, and the *F*-measure is close to 60% to more than 80%. The reason for the low recall precision was because our name lists were not exhaustive enough to cover the vast majority of the Kurdish names.

1998 ACM Subject Classification I.2.7 Natural Language Processing

Keywords and phrases Proper Noun Recognition, Named Entity Recognition, Information Extraction, Natural Language Processing, Kurdish

Digital Object Identifier 10.4230/OASIcs.SLATE.2017.19

1 Introduction

This paper suggests a method for proper noun recognition in Kurdish texts. Proper nouns are not capitalized in the vast majority of Kurdish texts. Despite recent intentions for using capitalization which are written in Latin script, this is technically not possible with Persian/Arabic script because it does not support capitalization formats. Also many proper nouns might have other grammatical forms such as being used as a verb, adjective, and object name. This causes word sense ambiguity in processes such as machine translation, information retrieval, and semantic analysis [4]. Kurdish is also considered a less-resourced [6, 14, 24, 5]. For example, the language does not have annotated corpora to be used as the required resources for most of Natural Language Processing (NLP) activities [4, 24]. We suggest a method that uses two name dictionaries, a gazetteer, a set of trigrams extracted from an untagged corpus, and a small set of hand-crafted rules.

Proper nouns recognition is part of Information Extraction (IE) and a subcategory of Named Entity Recognition (NER). NER is an application in NLP which extracts person names, locations, organizations, and generally, named entities from a text. This application is related to several other NLP and Computational Linguistics (CL) applications such as Machine Translation (MT) and Information Retrieval (IR).

As far as we are aware, there is no NER for Kurdish at the time of writing this paper. Sheykh Esmaili et al. [25] are the only scholars who have provided a significant research in Kurdish IR and have recognized several issues in this context [25]. In this research, among different categories of named entities, we have focused on person names detection.

6th Symposium on Languages, Applications and Technologies (SLATE 2017).
Editors: R. Queirós, M. Pinto, A. Simões, J. P. Leal, and M. J. Varanda; Article No. 19; pp. 19:1–19:13

Open Access Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Kurdish is a language that is mainly spoken by native Kurds in Iran, Iraq, Turkey, and Syria, alongside Kurdish communities in other countries such as Armenia, Lebanon, Egypt [6]. The population who speak the language is estimated about 30 million [11, 6]. Kurdish is multi-dialect from the Indo-European root [7]. Although different scholars have categorized its dialects differently, a considerable majority refer to it as Northern Kurdish (Kurmanji), Central Kurdish (Sorani), Southern Kurdish, Gorani, and Zazaki that include several sub-dialects [3, 6, 14]. Kurdish is written using four different scripts, which are modified Persian/Arabic, Latin, Yekgirtû(unified), and Cyrillic. The popularity of the scripts differ according to the geographical and geopolitical situations. Latin script uses a single character while Persian/Arabic and Yekgirtû in a few cases use two characters for one letter [6] (e.g., "ڕ" in Persian/Arabic and "sh" in Yekgirtû for "û" and "ş" in Latin, respectively). The Persian/Arabic script is even more complex with its RTL and concatenated writing style auto[6].

Although capitalization is possible in Latin scripts, this method neither has widely been practiced nor it has been standardized in Kurdish. However, there are evidence of attempts of practicing capitalization in Latin based texts, though it mainly covers Kurmanji texts but not in the texts written in Sorani (see [20, 30]). There are other languages such as Persian, Arabic, and Chinese that face this issue in a similar manner, for which a number of studies have been conducted, which would be helpful in the Kurdish proper noun computational study.

1.1 Kurdish Person Names

From the semantic perspective, Kurdish person names are intended to carry a meaning related to tangible objects such as waterfall, spring, and rain. This characteristics can be seen in many languages, though to a different extent. For instance, in Native American (indigenous American) languages this is a very popular case [27]. A large number of person names in Kurdish are Islamic (Arabic) nouns. However, the usage of different category of names, rooted in Kurdish history or culture has been growing since a few decades ago. In this approach to naming, historical names such as mythical or legendary Kurdish names, and the name of entities in nature are used as person names [1]. Many of these names have multiple linguistics categories, for example they can also appear as nouns or adjectives.

To illustrate, Table 1 shows several examples Kurdish names alongside their possible part-of-speech (POS) formats. We illustrate how these examples cause word sense ambiguity in MT and IR through a few examples. As the first example, "wenewşe hatwe.", might be interpreted as "The violet has come.", as a sign of spring, or "Violet has come.", meaning that a female whose name is Violet has just arrived. As another example, "çaw le baran ke!" could be interpreted as "Look at Baran!", which intends to ask for looking at a female whose name is "Baran". Also it can be interpreted as "Look at the rain!". Again, "înca baran degêrêtewe" can be interpreted as "then it protects against rain" or as "then Baran is telling the story".

Another issue is unusual homographs for names that are originally non-Kurdish, but have been transformed in a way that have made them a homograph to Kurdish words. For instance, "brayim" or "birayim" as a nickname for "îbrahîm", which is a homograph that might be interpreted as an Arabic name or might be interpreted as "I am her brother" or "I am his brother". Many Arabic names have been transformed into nicknames in the Kurdish context that sometimes officially appear in legal documents and ordinary texts as well. For example, "xule" for "xizir", "ebe" and "ewla" for "abudullah", and "bile" for "îbrahîm". This diversity can be seen for other foreign names such as English as well for which there is no standard that should be followed.

Table 1 Examples of Person Names in Kurdish – The first column shows the Kurdish name, the second shows the meaning of the name, and third shows the possible POS that the name might have in a sentence.

Name	Sense	Other POS formats
akam	result	Noun
amanj	goal	Noun
aso	horizon	Noun
azade	liberate, free	Noun, PP Verb
baran	rain	Noun
hawkar	colleague	Noun, Verb
sakar	simple, basic	Noun
tavge	waterfall	Noun
vareen	rain	Verb, Noun
wenewşe	violet	Noun, Adjective

Table 2 Examples of Arabic names in Kurdish – The first column shows the names in Arabic, the second shows the different orthographic appearance of the names, and the third shows different abbreviations of the names.

Arabic	Kurdish (orthographic formats)	Nickname
احمد	ehmed, ahmad	ehe, aha
حسين	husên, husen, hussên	wuze
محمد	mohammad, mohamed, muhemed, muhemmed	heme, mihe
قادر	kadir, qadir, ghader	kale, gale, ghalah
عثمان	othman, usman, osman	wetman

Table 2 shows a number of samples of Arabic names and some of their different formats in Kurdish alongside the nicknames that are used for the names. We have only showed the Latin version of Kurdish in this table. One can realize the issues that NER might face if one considers all these varieties.

Furthermore, another issue raises when many Arabic names that have multiple POS formats and their non-proper-name formats are also used in Kurdish. For instance, "جمال" which in Kurdish Latin script is written as "cemal" might mean "beauty" or "face" according to the context. In our approach, we have listed this kind of names in the Kurdish names dictionary in order to let the algorithm to apply the rules based on the words around the name (trigrams) and then make a decision about whether it should be taken as a proper name or not. These could have been done by adding features to a single dictionary and labeling each entity with appropriate category as Arabic, Kurdish, and such. However, we preferred to keep the lists in different dictionaries in this development stage. The way that this case is tackled might affect the performance of the devised algorithm either positively or negatively, but according to our project schedule and objectives, we did not implement more than one version of this application.

Among the other issues, we would like to address the problem with geographical names coming from the national or formal languages of the countries were Kurds located. For example, China and Austria are called 'sîn" and "nemsa" respectively, in Iraqi Kurdistan

that have been borrowed from Arabic. They are called "çîn" and "otrîş" in Iranian Kurdistan that are the way that these countries are called in Persian.

The rest of this paper is organized in the following sections. Section 2 reviews the related work. Section 3 discusses the methodology and presents the suggested method including an architecture, data collection steps, and a devised algorithm to implement the method. Section 4 presents the evaluation method and reports on the conducted experiments. Section 5 discusses the outcome of the experiments. Finally, Section 6 provides the conclusion and addresses the future work.

2 Related Work

Similar to other topics in NLP and CL, NER is significantly language and application specific. NER has been a focal point for many years and it has well-established architectures, practical approaches, and solutions for diverse applications in widely-studied languages such as English [10]. A large population of NLP and CL researchers worked on NER and have suggested practical approaches to the subject [15, 18, 26, 16, 2, 8].

Although all of these resources are helpful, most of them are working based on existence of proper computational resources. This is not the case in Kurdish NLP and CL, therefore, we review the works which have targeted the languages such as Chinese and Arabic which have similar proper nouns characteristics to Kurdish, for example, lack of capitalization. We also review the related work on languages such as Urdu which not only do not apply capitalization but also considered as less-resourced languages.

Sun et al. [26] provide an NER application based on statistical approach for Chinese [26]. Chinese NER faces another problem which is lack of space for marking word boundaries. However, researchers of this work rely on manually tagged data sets large enough for statistical/probabilistic approach, which is currently not applicable for Kurdish.

Tsai et al. [29, 28] report on "Mencius" an NER for Chinese in which they have used a hybrid model by combining rule-based and Machine Learning (ML) based approaches [28, 29]. They perform three experiments one of which is a rule-based method which uses name list and gazetteers with 32000 entries. The second one is an ML based experiment and the third is a hybrid approach that combines the two previous methods. The results show different level of accuracy for different type of entities. The ML approach is not applicable in the current situation of Kurdish hence we are interested in the first method of this study which is applicable in the absence of an established Language Model (LM) and corpora.

Shaalan and Raza [22] have developed a proper noun recognizer for Arabic using rule-based methods. They have suggested an architecture that has three major blocks. A gazetteer that includes several dictionaries, a grammar configuration that recognizes person names using regular expression patterns, and a filtration mechanism to reject invalid person names. They have reported that their system achieved 85.5% for the precision and 89% for recall measures [22].They have expanded and slightly modified their work in [23]. A recent survey on Arabic NER, [21] reports that both rule-based and ML based approaches have been successful hence a hybrid method has been suggested that utilizes the advantaged of each method in a single one.

Riaz has suggested a rule-based NER for Urdu [19]. The author provides a list of challenges that Urdu NER faces, several of which are common with Kurdish such as capitalization, ambiguity, spelling variations, loan words, and resource challenges. The suggested method uses the hand-crafted rules to form a Finite State Automata (FSA) based on lexical indications. The rules are categorized as corpus-based, heuristic-based, and grammar-based.

The approach also utilizes a 6-gram that have been extracted form available Urdu corpora. The paper reports that the results showed promising performance when the method is compared with other NER methods.

3 Methodology

The "practical architecture"s which have been applied by the mentioned researchers in Section 2 are not practical in our research for the lack of required resources. As a result, we have developed a revised architecture that fits the current situation of Kurdish NLP and CL which is in its infancy.

Our approach to the person name recognition is benefited from the work by Shaalan and Raza that suggest a rule-based person name recognition in Arabic [22]. It is also based on the work by Tsai et al. who proposes a hybrid model by combining rule-based and Machine Learning [28]. Furthermore, we consider the work by Riaz which provides a rule-based NER for Urdu [19].

However, our approach differs, in different ways, from what have been proposed in the mentioned studies. First, in our architecture the arrangement and type of dictionaries are different. This rearrangement and categorization enables the method to accept other types than person names and also simplifies the arrangement of dictionaries. For example, having locations as a separate list, allows the application to be expanded to recognize the location in the future. In fact, because no annotated corpus currently exists for Kurdish, we cannot label the entity types by using probabilistic methods. Therefore, in our architecture we have separated the name lists. Obviously, the hand-crafted rules also differ from the suggested methods because of these rules are mainly language-specif. In addition, we have devised and presented an algorithm that shows the implementation configuration.

3.1 Architecture

Figure 1 shows the proposed architecture for Kurdish proper nouns recognition.

Below the components of the architecture are explained.

- A Gazetteer- This includes a subset of proper names as it is used in most of more established NER approaches.
- A dictionary of Kurdish names- It includes person names that mainly have Kurdish origins.
- A dictionary of Arabic names- It includes Arabic names that can certainly be identified in Kurdish texts. The nickname/moniker equivalents are kept in item 2 (The dictionary of Kurdish names).
- A set of rules- The rules are hand-crafted and include words that could appear before or after a proper noun allowing the candidate names to be taken as proper nouns with a higher probability. For example, if any member of the subset ("aga", "axa", "beg", "xatûn", "xanim") appears after a candidate name, the candidate name is considered as a found name.

3.2 Data Collection

We collected data in several steps and organized them in different data containers as we will present them below.

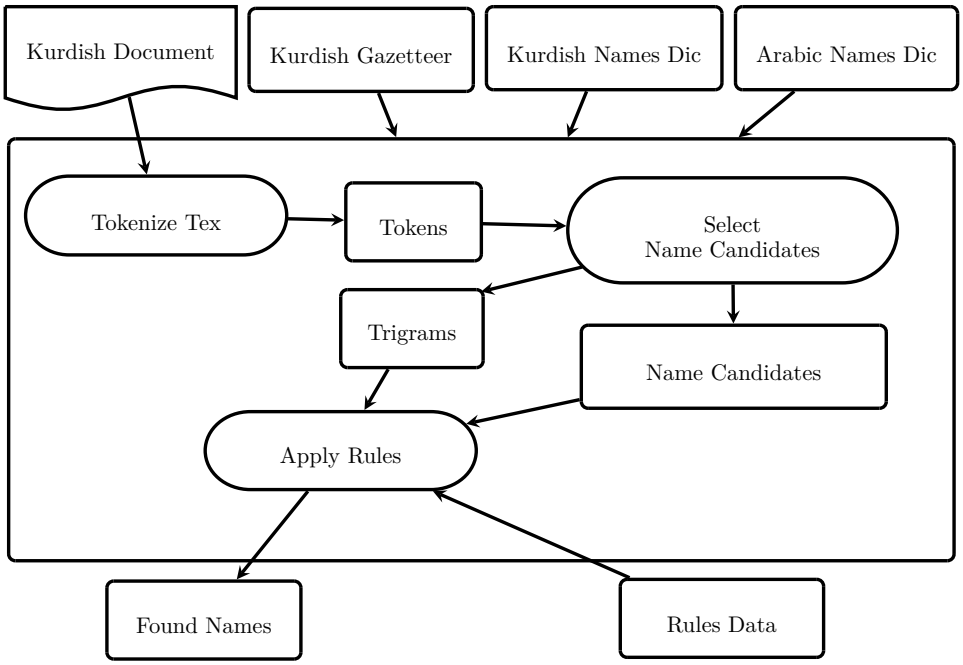

Figure 1 Kurdish PNR Architecture - The architecture is based on three dictionaries and a set of rules which is used by the internal part of the system to find out the Proper Names.

3.2.1 Gazetteer

The gazetteer was manually created. It includes geographical names and famous person names including a number of well-known people such as poets, philosophers, and leaders. It is a developing list and was not intended to be complete for this research because the aim of this research is to recognize person names. The data was collected from different online resources such as GeoNames[1] and wherever needed the transliterated and their diacritics were unified.

However, at this stage we have not intended to recognize location entities hence we have not focused on data collection for this item. We will address this case again when we discuss the future work in Section 6.

3.2.2 Kurdish Names Dictionary

The data for Kurdish names dictionary was collected from various available online resources. The names written in Persian/Arabic were transliterated to Latin and transformed into their lower case before saved. Also the names that were not Kurdish or Arabic, for example, Assyrian and Chaldean names were inserted into this dictionary at this version of Kurdish PNR.

Because diacritics are not unified in Kurdish texts, for example, “ ‘ ” is used instead of “ˆ” in some texts, therefore, the diacritics were also unified (harmonized) according to the more accepted version, which is “ˆ”. Afterwards, duplicates were removed from the list and the list alphabetically sorted and it was written into a text file. Finally, although Arabic

[1] Available from `http://www.geonames.org/`.

names are popular in Kurdish communities, we have separated these names from the original Kurdish names.

We used [13, 12, 9] for the name lists construction. We also used name entries from the list of accepted students to the universities of Iraqi Kurdistan region, which have been announced by Ministry of Higher Eduction and Scientific Research (Kurdistan Regional Government) [17]. Although this latter case could have improved our lists in terms of the number of entries, for various technical reasons we were not able to automatically convert the list to the format that our application could handle, therefore, we manually processed the list and added some of the names that we did not have in our names lists.

3.2.3 Arabic Names Dictionary

Many Arabic names are popular in Kurdish communities. There are several historical reasons for this phenomenon, the discussion of which is beyond the scope of this dissertation. However, since several decades ago the usage of Kurdish names has significantly increased. Nevertheless, Arabic names are quite popular and hence they appear in many texts very frequently. This dictionary includes these names.

Despite the popularity of Arabic names in Kurdish, they do not follow the same spelling and orthography as they are used in Arabic. Also there is no orthographic standard for writing these names, hence there might be several versions of a single name no matter if Latin or Persian/Arabic script has been used (see Table 2).

We created Arabic names list based on our knowledge with regard to the Kurdish culture and familiarity with both Arabic and Kurdish languages.

3.2.4 Rules Set

The rule set consists of four lists as below:

- List of salutations that appear before a proper noun such as “mamosta” that is used to call a teacher or lecturer or a Muslim cleric or “kak’ that is used to friendly and respectfully call a male person, literally meaning “big-brother”.
- List of salutations that appear after a proper noun such “axa” or “xan”, which appear after a male and female person, literally meaning “sir” and “madam”. The second salutation is used for male person in Iranian Kurdistan.
- List of words that appear before a proper noun.
- List of words that appear after a proper noun.

We hand-crafted these rules based on our familiarity with Kurdish. We do not claim that this rules are complete rather we have suggested them to show that our approach is practical.

3.3 Algorithm

An algorithm has been developed in order to recognize proper nouns in Kurdish texts. The algorithm has been presented below. It reads the required dictionaries as its knowledge base, defines a number of lists to hold the candidate names, trigrams, and found names, and reads the input text. The input text will be tokenized and then processed by matching each token with the dictionaries. The algorithm is able to apply a stemming process, by calling a stemmer, on the token under investigation if it did not find it in the name lists. The stemmer acts as a suffix stripper in this case. Similar to many other languages, proper nouns can have suffixes in Kurdish. When a proper noun has a suffix, it cannot be found in

name dictionaries. We need a stemmer that is able to strip the suffix of the token in order to produce a proper searching item for the name dictionaries. For example, in the sentences "baran**im** dît." if we strip the bold letters, we might have a person name. As another example, in the expression "le kurdistan**ê**" if we strip the bold letter, we will definitely have a proper noun. The proper nouns are shown by underlining them in both mentioned cases.

Algorithm 1 Kurdish Person Name Recognition.

```
function findKurdishPersonName(inputText)
    nameCandidates ← null
    nameCandidatesRuleApplied ← null
    trigramsAll ← null
    trigramsRuleApplied ← null
    isNameCandidate ← false
    isCheckedCandidate ← false

    read KurdishNames, KurdishGazetteer, ArabicNames, Rules
      ▷ Notice: The order of evaluation is important. It should be ordered according to the number
        of entries in each list.
    for token in inputText do

        if token is in KurdishGazetteer or token is in ArabicNames then
            isCheckedCandidate ← true
        ▷ Notice: If the token is not found as it is, stem the token and compare the result.
        else if token is in KurdishNames or stemmed-token is in KurdishNames then
            isNameCandidate ← true
            if Rules Apply or (either predecessor-token or successor-token) is in
                                (KurdishGazetteer or ArabicNames or KurdishNames) then
                isCheckedCandidate ← true
            end if
        end if

        if isCheckedCandidate then
            append token to nameCandidates
            append token to nameCandidatesRuleApplied
            trigram ← concat(predecessor, token, successor)
            append trigram to trigramsAll
            append trigram to trigramsRuleApplied
            isCheckedCandidate ← flase
        else if isNameCandidate then
            append token to nameCandidates
            trigram ← concat (predecessor, token, successor)
            append trigram to trigramsAll
            isCheckeCandidate ← flase
        end if

    end for

    return nameCandidates, nameCandidatesRulesApplied, trigramsAll, trigramsRulesApplied
end function
```

4 Evaluation

We used 15 documents of different sizes to test the accuracy of the method, ranging from several hundred to several thousand tokens, of which 8 documents were in Kurmanji and 7 were in Sorani. We ran the suggested algorithm once without and once with the stemming process on each document and saved the results separately. We also manually extracted the person names from each input text. The outputs of the algorithm, the suggested names, were compared against the manually extracted names whereby we calculated the Precision and Recall parameters and the F-measure for each document.

Table 3 Kurdish PNR in Kurmanji Texts. TP: True Positive, FP: False Positive, and FN: False Negative were extracted by examining the results against the texts. P: Precision, R: Recall, and F: *F*-measure were calculated using Equations 2, 3, and 1 respectively. The last row shows the average performance of the experiments.

No	Size (Words)	**TP**	**FP**	**FN**	**P**	**R**	***F***
1	1605	7	0	3	1	0.7	0.82
2	1798	4	0	6	1	0.4	0.57
3	1910	11	1	6	0.92	0.65	0.76
4	2200	8	0	6	1	0.57	0.73
5	2300	10	0	9	1	0.53	0.69
6	2112	11	1	6	0.92	0.65	0.76
7	2520	9	0	5	1	0.64	0.78
8	2400	11	2	7	0.85	0.61	0.71
				Average	0.96	0.59	0.73

The classic evaluation method for NER is based on *Precision, Recall*, and F_1 measure as shown in Eq. (1). In our case, *Recall* is the ratio of the number of correctly found names to the total names in the sample. *Precision* is the ratio of the number of correctly found names to the total found names.

$$F_1 = \frac{2PR}{P+R} \tag{1}$$

P and R are calculated by the following formulas:

$$P = \frac{T_P}{T_P + F_P} \tag{2}$$

$$R = \frac{T_P}{T_P + F_N} \tag{3}$$

In Eq. (1), P stands for Precision and R for Recall. In Equations (2) and (3), T_P stands for *True Positive*s, which in our case are correctly recognized Person Names; in Eq. (2), F_P stands for *False Positive*s, which are wrongly recognized Person Names; and in Eq. (3) F_N stands for *False Negative*s, which are unrecognized Person Names.

Table 3 shows the result of experiments for Kurmanji texts. Table 4 shows the result of experiments for Sorani texts.

Currently there is no NER for Kurdish, hence we compared the results with the outcome of the works which we have addressed in Section 2. Shaalan and Raza report a *Precision* between 84.2% to 94%, a *Recall* between 84.7% to 96.8%, and an *F-measure* between 84.4% to 95.1% [22]. Riaz reports a *Precision* of 91.5%, a *Recall* of 90.7%, and an *F-measure* of 91.1% for one data set and an *F-measure* between 72.4% to 81.6% for another one [19]. Our method shows a *Precision* between 80% to 100%, a *Recall* between 40% to 77%, and an *F-measure* between 57% to 87%. Close investigations of the results, by looking into the manually extracted names, showed that the majority of unrecognized Person Names were non-Kurdish names.

5 Discussion

As Tables 3 and 4 show, the proposed architecture and the devised algorithm work with a high precision in most of the cases tested. However, they also show that the recall ratio is

Table 4 Kurdish PNR in Sorani Texts. TP: True Positive, FP: False Positive, and FN: False Negative were extracted by examining the results against the texts. P: Precision, R: Recall, and F: *F*-measure were calculated using Equations (2), (3), and (1) respectively. The last row shows the average performance of the experiments.

No	Size (Words)	**TP**	**FP**	**FN**	**P**	**R**	***F***
1	1805	7	0	3	1	0.7	0.82
2	1994	4	0	6	1	0.4	0.57
3	2507	12	0	7	1	0.63	0.77
4	2302	10	0	3	1	0.77	0.87
5	2105	10	0	3	1	0.77	0.87
6	2900	11	1	4	0.92	0.73	0.81
7	2320	8	2	3	0.8	0.73	0.76
			Average		0.96	0.68	0.78

very low in several cases and low in majority of them. The values show that the method is performing in the acceptable boundary providing a better upper-bound for *Precision* when it is compared to [22] and [19]. However, its *Recall* does not perform as its *Precision* which causes the *F-measure* to have a low upper-bound. This is because the name lists, the dictionaries, should be expanded to cover more names. Our dictionaries mainly cover popular names in Iraqi Kurdistan, while person names in other Kurdish speaking regions have been influenced by other countries culture such as Iranian, Turkish, and Syrian culture. Moreover, the same name might be spelled differently, which means that all formats of the names must be found and recorded. < This shows that although the proposed method works properly, the underlying data that are based on the fundamental components of the proposed architecture in Section 3 must be expanded and enriched more in order to cover foreign names as well. The investigation also showed that a number of unrecognized names has been captured and written in different format than what could have been found in the supporting dictionaries.

Importantly, we assessed the case of stemming and trigrams. The results revealed stemming has a significant impact on recall and trigrams considerably affect the precision. Figures 2 and 3 below show these findings for Kurmanji and Sorani dialects respectively. Differences between Kurmanji and Sorani results are mainly coming from two aspects. The first aspect is the usage of Turkish names in the Kurdish community in Turkey and the second is the stemming accuracy level. For the Sorani dialect the usage of other Iranian names, such as those that are mainly considered as Persian names affects texts that are written based on the Iranian sub-dialect of Sorani. However, whether increasing the order of n-grams to a higher order, for example four or five, leads to an improved recall or precision is yet to be investigated.

Finally, it is worth mentioning that although we have incorporated gazetteer into the architecture, geographical and place names have not been our target, hence their recognition has not participated in the evaluation process.

6 Conclusion

In this paper we presented the challenges that exist in Kurdish Named Entity Recognition. The research focused on Person names recognition. The result showed that the presented

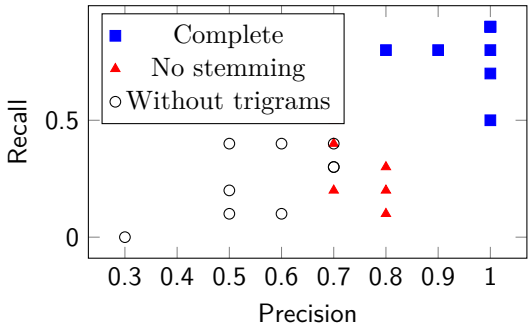

Figure 2 The Impact of Stemming and Trigrams on Precision-Recall (Kurmanji).

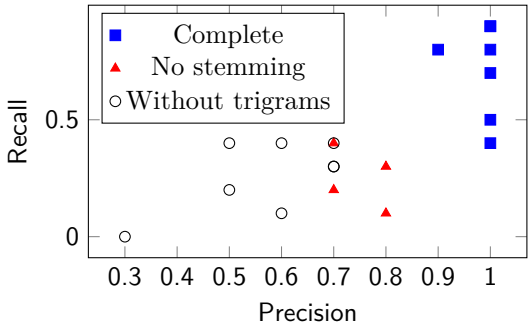

Figure 3 The Impact of Stemming and Trigrams on Precision-Recall (Sorani).

method performed well in the absence of underlying data that is necessary for other well-established approaches to NER. The precision of the method is between 80% to 100% with the average more than 95%, the recall is between 40% to 80% with the average of more than 60%, and the F-measure is close to 60% to more than 80%.

6.1 Future Work

There are areas that should be studied further with regard to PNR, and in a general sense NER, in Kurdish. Most importantly, the method should be improved to be able to recognize not only person names but also entities in the general sense such as locations and organizations. For the geographical names, we expect that the expansion of Gazetteer and enhancing the rule-set would do the purpose. However, for the entities of organization type the case must be properly investigated. Moreover, the gender categorization should be implemented in person names recognition. Also the efficiency of trigrams should be tested and compared to other orders of n-grams in order to find the optimum format. Furthermore, the multi-segment names should also be considered and recognized properly. In addition, the dictionaries must be expanded to cover names from other Kurdish speaking regions in other countries. Similarly, the dictionaries should be augmented by different spellings of names.

Equally important, the method should be revisited when the Kurdish NLP and CL are matured enough by having proper language models that allow researchers to apply probabilistic approaches and ML-based NER. A comparison of combination of the methods, alongside the application of each method independently, could be the subject of another series of research.

Acknowledgements. We would like to express our warm appreciations to Dr. Dzejla Medjedovic an Assistant Professor and Vice Dean of Graduate Program at the University Sarajevo School of Science and Technology (SSST) for reviewing this paper and providing influential recommendations. We would also like to thank the anonymous reviewers for their constructive suggestions.

References

1 Lazgin Al-Barany, Asma Albamarni, and Dilggash M. Shareef. Kurdish personal names in Kurdistan of Iraq: A sociolinguistic perspective, 2014. https://www.academia.edu/9662401/Kurdish_Personal_Names_in_Kurdistan_of_Iraq_A_Sociolinguistic_Perspective.

2 Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning*, volume 7, pages 708–716, 2007.

3 Geoffrey Haig and Ergin Öpengin. Introduction to special issue-Kurdish: A critical research overview. *Kurdish Studies*, 2(2):99–122, 2014.

4 Hossein Hassani. Kurdish interdialect machine translation. In *VarDial Workshop*, pages 63–72, April 2017.

5 Hossein Hassani and Rahel Kareem. Kurdish text to speech (KTTS). In *Tenth International Workshop on Internationalisation of Products and Systems*, pages 79–89, 2011.

6 Hossein Hassani and Dzejla Medjedovic. Automatic Kurdish dialects identification. *Computer Science & Information Technology*, 6(2):61–78, 2016.

7 Amir Hassanpour. *Nationalism and language in Kurdistan, 1918-1985*. Edwin Mellen Pr, 1992.

8 Ulf Hermjakob, Kevin Knight, and Hal Daumé III. Name translation in statistical machine translation-learning when to transliterate. In *Association for Computational Linguistics*, pages 389–397, 2008.

9 Hesami. Kurdish definition, origin and usage of names, 2016. http://www.hesami.com/names/kurdish/.

10 Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall, 2 edition, 2008.

11 Kurdish Academy of Languages. The Kurdish Population, 2016. http://www.kurdishacademy.org/?q=node/199.

12 Kurdish Daily. Kurdish names for your baby, 2016. http://ekurd.net/mismas/kurdishnames.htm.

13 Kurdish Institute of Paris. Kurdish Names, 2016. http://www.institutkurde.org/en/kurdorama/kurdish_baby_names.php.

14 Shervin Malmasi. Subdialectal differences in Sorani Kurdish. In *Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 89–96, 2016.

15 Inderjeet Mani, T Richard MacMillan, Susann Luperfoy, Elaine Lusher, and Sharon Laskowski. Identifying unknown proper names in newswire text. In *Workshop on Acquisition of Lexical Knowledge from Text*, pages 44–54, 1993.

16 Gideon S. Mann and David Yarowsky. Unsupervised personal name disambiguation. In *Seventh conference on Natural language learning at HLT-NAACL*, volume 4, pages 33–40, 2003.

17 Minstray of Higher Education and Scientific Research. Admitted students in 2010, 2016. http://www.mhe-krg.org/ku/node/698.

18 Thierry Poibeau and Leila Kosseim. Proper name extraction from non-journalistic texts. *Language and Computers*, 37(1):144–157, 2001.

19 Kashif Riaz. Rule-based named entity recognition in Urdu. In *Named Entities Workshop*, pages 126–135, 2010.

20 Rudaw, 2015. `http://rudaw.net/sorani`.

21 Khaled Shaalan. A survey of arabic named entity recognition and classification. *Computational Linguistics*, 40(2):469–510, 2014.

22 Khaled Shaalan and Hafsa Raza. Person name entity recognition for Arabic. In *Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 17–24, 2007.

23 Khaled Shaalan and Hafsa Raza. NERA: Named entity recognition for arabic. *Journal of the American Society for Information Science and Technology*, 60(8):1652–1663, 2009.

24 Kyumars Sheykh Esmaili. Challenges in Kurdish text processing. *arXiv preprint arXiv:1212.0074*, 2012.

25 Kyumars Sheykh Esmaili, Shahin Salavati, and Anwitaman Datta. Towards Kurdish information retrieval. *Transactions on Asian Language Information Processing (TALIP)*, 13(2):7, 2014.

26 Jian Sun, Jianfeng Gao, Lei Zhang, Ming Zhou, and Changning Huang. Chinese named entity identification using class-based language model. In *19th International Conference on Computational Linguistics*, pages 1–7, 2002.

27 Tribal Directory. American Indian Names, 2016. `http://tribaldirectory.com/information/american-indian-names.html`.

28 Tzong-Han Tsai, Shih-Hung Wu, and Wen-Lian Hsu. Mencius: A Chinese named entity recognizer using hybrid model. In *Research on Computational Linguistics Conference XV*, pages 193–209, 2003.

29 Tzong-Han Tsai, Shih-Hung Wu, Cheng-Wei Lee, Cheng-Wei Shih, and Wen-Lian Hsu. Mencius: A chinese named entity recognizer using the maximum entropy-based hybrid model. *International Journal of Computational Linguistics and Chinese Language Processing*, 9(1), 2004.

30 Wikipedia. Hûn bi xêr hatin Wîkîpediyaya kurdî, 2016. `https://ku.wikipedia.org/wiki/Destp%C3%AAk`.