Kurdish Interdialect Machine Translation

Hossein Hassani
University of Kurdistan Hewlêr
Sarajevo School of Science and Technology
hosseinh@ukh.edu.krd
hossein.hassani@stu.ssst.edu.ba

Abstract

This research suggests a method for machine translation among two Kurdish dialects. We chose the two widely spoken dialects, Kurmanji and Sorani, which are considered to be mutually unintelligible. Also, despite being spoken by about 30 million people in different countries, Kurdish is among less-resourced languages. The research used *bi-dialectal* dictionaries and showed that the lack of parallel corpora is not a major obstacle in machine translation between the two dialects. The experiments showed that the machine translated texts are comprehensible to those who do not speak the dialect. The research is the first attempt for inter-dialect machine translation in Kurdish and particularly could help in making online texts in one dialect comprehensible to those who only speak the target dialect. The results showed that the translated texts are in 71% and 79% cases rated as *understandable* for Kurmanji and Sorani respectively. They are rated as *slightly understandable* in 29% cases for Kurmanji and 21% for Sorani.

1 Introduction

This paper discusses Intralanguage Machine Translation (IMT) among Kurdish dialects. The two most widely spoken Kurdish dialects are Kurmanji and Sorani which are considered to be mutually unintelligible (Hassanpour, 1992). Furthermore, the language is among less-resourced languages (Sheykh Esmaili, 2012; Sheykh Esmaili et al., 2014). However, this research shows that, in the absence of large parallel corpora, a word-for-word translation approach based on a bidialectal dictionary provides a reasonable translation output between the dialects. This improves mutual intelligibility among Kurmanji and Sorani users in the online textual environment.

Our aim is to show that lack of corpus is not a major obstacle for providing an inter-dialect (intralingual) machine translation between Sorani and Kurmanji. Our method intends to transfer the general meaning of texts in online media in one dialect to those audience who speak the other. To that extent, the output is not considered to be a literary translation nor it is able to transfer all grammatic features of the source to the target dialect.

Machine Translation (MT) is primarily understood as using computers for translating a language into another, or in other words, as automated *inter-language* translation. The main motive of MT is to make a language L_1 intelligible to whom who do not speak it by presenting it in a language L_2, which might be the audiences' own language or a language which they are able to understand. However, there are several languages such as Chinese, Arabic, and Kurdish that encompass several dialects which are mutually unintelligible (Tang et al., 2008; Farghaly and Shaalan, 2009; Sadat et al., 2014). In this respect, the translation between the dialects are of the *intralanguage* nature rather than *interlanguage*.

Kurdish is the name given to a number of distinct dialects of a language spoken in the geographical area touching on Iran, Iraq, Turkey, and Syria. However, Kurds have lived in other countries such as Armenia, Lebanon, Egypt, and some other countries since several hundred years ago. The population who speak the language is estimated about 30 million (Kurdish Academy of Languages, 2016; Hassani and Medjedovic, 2016).

Dialect diversity is an important characteristic of Kurdish. This diversity, the name of dialects, and their geographic distribution have been

Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 63–72,
Valencia, Spain, April 3, 2017.

of interest for linguists who have been studying Kurdish. Kurdish is multi-dialect from the Indo-European root (Hassanpour, 1992). 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 (Haig and Öpengin, 2014; Hassani and Medjedovic, 2016; Malmasi, 2016). The populations that speak different dialects of the language differ significantly. The majority of Kurmanji speakers are located in different countries, such as Turkey, Syria, Iraq, Iran, Armenia, Lebanon, just to name the mainlands. The second popular dialect is Sorani, which is mainly spoken among Kurds in Iran and Iraq. Zazaki is spoken in Turkey. Gorani is primarily spoken in Iran and Iraq (Izady, 1992; Hassanpour, 1992). 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. The Persian/Arabic script is even more complex with its RTL and concatenated writing style auto(of Language, 2016).

We are facing the "knowledge acquisition bottleneck", which basically occurs in the early stages of Natural Language Processing (NLP) and Computational Linguistics (CL) studies (Schubert, 2015), hence we are interested in investigating of interdialect Kurdish translation in the absence of parallel corpora. Our hypothesis is that despite the mutual unintelligibility between the two dialects, a word-for-word translation would be able to transfer the core meaning of texts in one dialect into the other. To illustrate, Ballesteros and Croft (1996) have reported on the applicability using dictionaries in certain situations such as Information Retrieval (IR). This solution can be used while the necessary background knowledge is prepared for statistical MT.

The remainder of the article is organized in the following sections. Section 2 reviews the literature. Section 3 provide the method that is used in developing an IMT for Sorani-Kurmanji. Section 4 presents the performed experiments on the developed IMT and evaluates the results. Section 5 discusses the findings and the outcome of the experiments and analyzes the results. Finally, section 6 summarizes the findings, provides the conclusion, and addresses the future work.

2 Related Work

Zhang (1998) discusses inter-dialect MT between Cantonese and Mandarin as the two most important Chinese varieties, which are considered to be mutually unintelligible. Zhang (1998) discusses the differences between the two dialects at the level of sound systems, grammar rules, and vocabulary, based on which a method for inter-dialect MT between the two dialects has been provided. Zhang (1998) suggests that as the dialects of a language usually share a common standard written form, the target of inter-dialect MT is better to be the spoken dialects. The method has been implemented by using a Word collocation list, a Mandarin-Cantonese dictionary and a handful number of rules to handle syntactic differences. Zhang (1998) addresses the immediate purposes of the developed systems as to facilitate language communication and to help Hong Kong students to write standard Mandarin Chinese. However, he has not reported on the evaluation of the system and the level of intelligibility of the system's output by the targeted audience. Furthermore, the research reports that a Mandarin-Cantonese corpus has been built, but it does not mention how it has been created nor how it has been used in inter-dialect MT. Moreover, although it has been mentioned that the rules for the syntactic difference between the dialects are applied based on a knowledge base, it is not clear whether this knowledge base uses a Part-of-Speech (POS) tagger or an annotated corpus or it has applied another approach.

Peradin et al. (2014) suggest a shallow-transfer rule-based machine translation for Western group of South Slavic language using Apertium platform which is a modular machine translation system. Peradin et al. (2014) have used morphological lexicons available on Apertium repository.

Nakov and Tiedemann (2012) worked on Macedonian-Bulgarian machine translation as *close languages*. They have put their assumption based on the morphological and lexical similarities and have used statistical approach combined with word-for-word translation to show that MT is possible without having large corpora. Although the work technically could be of help for inter-dialect MT, it is not an IMT study in principle.

Our search for finding more work on automatic translation among dialects, which we called IMT, did not yield any other significant work beyond what has been done by Zhang (1998). To illustrate, we refer to a recent publication, a comprehensive handbook by Chan (2014), which covers different aspects of MT and MT technologies. However, although the book addresses the MT status with regard to different languages, for the work related to inter-dialect MT it only refers to the studies by Zhang (1998).

As another evidence for lack of noticeable study on IMT we refer to "The first workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects" was conducted in 2014 (Zampieri et al., 2014) and consequently a "Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects" (Ass, 2015) took place in 2015. Although these events fairly covered several areas of close languages, none of the papers discussed IMT.

However, the literature on mutual intelligibility has a longer background of scholarly work and is also closely related to our research area in a broad sense (Voegelin and Harris, 1951; Pierce, 1952; Yamagiwa, 1967).

Cheng (1997) has measured the relationship among dialects of Chinese . Also Szeto (2000) tested the intelligibility of Chinese using a tape-recorded text (RTT), asking the participant group members to write down the recognized vocabularies (Szeto, 2000). In a slightly recent attempt, Kluge (2006) suggested some improvements with regard to the question-answering approach of standard RTT. However, in none of theses studies the computational aspects of the process have been of concern to the researchers.

Tang and van Heuven (2009) have performed an experiment on Chinese and assessed intelligibility among a number of its dialects. They discuss the adequacy of mutual intelligibility testing "to determine how different two languages or language varieties are". Their method is based on speech recognition at both word and sentence intelligibility level.

Munday (2009) refers to intralingual translation as "rewording" and describes it as the process of summarizing or rewriting a text in the same language. However, the majority of this work has focused on interlingual, particularly bilingual translation.

From a different perspective, Beijering et al. (2008) have studied the dialectal and inter-language intelligibility and perceived linguistic distance among Scandinavian dialects using Levenshtein algorithm. According to Beijering et al. (2008) the Levenshtein algorithm is able to successfully predict intelligibility among different languages/dialects.

3 Methodology

We were not able to apply the probabilistic approach in inter-dialect machine translation because of the lack of required infrastructure in terms of parallel, annotated, and tagged corpora at the time of conducting this project. Therefore, we aimed to use a method for *intralingual* (inter-dialect) MT between Sorani and Kurmanji that is applicable in the absence of large data. As a result, we used a modified version of the method suggested by Zhang (1998) in which a word collocation list, a bidialectal dictionary and a series of rules to handle syntactic differences between the dialects are used to perform inter-dialect MT between Mandarin and Cantonese. In our adaptation, we have not considered the grammatic differences of Kurmanji. There are two reasons for this, first, lack of the required resources such as tagged corpora which does not allow us to implement an efficient syntactic analysis, and second, the regional variations in Kurmanji (Öpengin and Haig, 2014) makes the rules more complicated.

Therefore, we have based our method on the development of the two *bidialectal* dictionaries, one for Sorani to Kurmanji equivalents, and the other for Kuramnji to Sorani. We have implemented a word-for-word translation, which is also known as word-by-word (in a number of texts it is also called literal) or direct translation. This is an incremental transformation of the source-language text into a target-language text without having any knowledge about phrasing or grammatical structure in the source or target language (Jurafsky and Martin, 2008).

3.1 Dictionary Development

We used web data, mainly websites of Kurdish media and universities in Iraqi Kurdistan region, for our data collection. In terms of the genre, we selected the texts that were about art, literature, sport, and education. The reason was that we were interested in assessing the efficiency/adequacy of

our method in helping Kurmanji users to be able to comprehend the online texts of ordinary day-to-day social genres written in Sorani and vice versa. We transliterated the texts in Persian/Arabic to Latin. We processed these texts and extracted their lexicon. We then used several Kurdish dictionaries to set the first sets of word equivalents, the lexicon, in the target dialect. For this purpose, we used (Demîrhan, 2007; Wîkîferheng and Ferhenga azad (*Azad Dictionary*), 2015; Ronahî, 2015; Mohammed Ali, 2008). The first three items are available online and the last one is in printed format. We also used our knowledge about the dialects and consulted language informants in the cases that the dictionaries could not resolve. This round of dictionary development process produced 6792 words out of which 2632 words are in Kurmanji and 4160 in Sorani.

3.2 Evaluation Method

We evaluated the efficiency of the implemented IMT by adapting the human raters method, which uses human experts to rate the translated texts. In this method several parameters are used such as fidelity or accuracy, intelligibility or clarity, and style (Fiederer and O´Brien, 2009; Ahsan et al., 2010). Although automated methods such as BLEU (bilingual evaluation understudy) (Papineni et al., 2002) have been implemented for MT evaluation, they perform more efficiently in the presence of proper corpora and language models, which were not available in our case.

We followed a combination of qualitative/quantitative approach for the evaluation process. In our adaptation of human raters, the translated texts are given to several speakers whose main dialects are not the same as the original text. Also some speakers will be chosen who have learned one of theses dialects as a second language and they do not have any familiarity, or at least any considerable familiarity, with the other dialect. The method then quantitatively evaluates the comprehensibility/understandability degree of the translated texts using this parameter. We also conduct a short interview with the human raters after they rated the text to qualitatively assess the result. We have not considered "style" parameter of human rating in our experiment because we have not evaluated the syntactic/parsing aspect of the translated texts.

For comprehensibility evaluation, Fiederer and O´Brien (2009) suggest 4 levels while Ahsan et al. (2010) suggest 5 levels. In our approach the translated texts is ranked in 5 categories: *not understandable*, *slightly understandable*, *understandable*, and *completely understandable*. Nevertheless, as it was mentioned, the translation evaluation is a subjective process no matter how one tries to quantify it. We briefly explain to the participants that they should rate a text as *not understandable* if they find that they cannot comprehend what the text is about; as *slightly understandable* if they know the meaning of a number of words and even a sentence but they do not have an overall comprehension of the text; as *understandable* if they comprehend the text but they do not know a few words; as *completely understandable* if they comprehend the text and they know the meaning of all words.

After rating process by each participant, we hold an interview to verify their understanding level. In this interview, we ask participants to tell us what the passages were about in their own words. Furthermore, if the given rate is *slightly understandable* or *understandable*, we ask the participant to explain what is the reason that they have not rated the text as *completely understandable*.

We also ask the participants to rank themselves with regard to their fluency in reading Kurdish texts in Latin, because many Kurdish speaking people in Iraqi Kurdistan either cannot read Kurdish texts in Latin or they are not fluent in reading this script [1]. We will ask more precise questions in the interview to verify the comprehensibility of the text, if a participant rates their Latin reading fluency below *good*.

Importantly, this study has not intended to consider the aesthetic aspects of the translation as an art, rather its intention is to evaluate the adequacy of this approach in the absence of large data that could support a high quality translation.

4 Experiments

We developed a transliterator to transliterate texts in Persian/Arabic script into Latin script. The dictionaries created in Latin script and where it was necessary the vocabulary was transliterated from Persian/Arabic into Latin. We manually removed

[1] In fact, most of the Kurdish speaking people, even if they are well-educated, might not be fluent in reading Kurdish texts either in Latin or Persian/Arabic or Cyrillic depending on the region who they live or have grown up.

Yekgirtû double-sign letters, such as “sh” in Yekgirtû and replaced with “ş” in Latin, when the source was created using Yekgirtû alphabet. We also uniformed the diacritics, for example by replacing “ˆ” with “^”.

For the evaluation of the implemented IMT, we arbitrarily chose 3 passages from Kurdish media. For the Sorani texts we transliterated the texts from Persian/Arabic script into Latin script. We then machine translated the texts and printed the output with the intelligibility rating printed alongside each translated text in the way that we mentioned in Section 3. We gave the texts to our participants in the human rating process. Out of 11 participants who were all native Kurdish speakers, 3 could only speak in Kurmanji, 5 could only speak in Sorani, 2 could speak in both dialects. There was one participant who was not native Kurdish speaker and has learned basic Sorani dialect. Except the latter case, the other participant ranked their fluency in reading Latin texts at least *very good*. Although one participant rated their Latin reading fluency as *beginner*, we verified the evaluation in the interview and found it appropriate. During the evaluation process, we did not explain to the participants that the texts were a machine translated results and only asked them to rate the text based on their understanding. We did not intervene or help in any case until the rating was finished.

The interviews showed that in all cases the rating was almost conforming with what had been assigned. However, when we asked participants who had rated a text *understandable* about why they did not find it *completely understandable*, we realized this was coming from the grammatical issues of the translated text. The participants replied that they had found the text not fluent from grammatical perspective.

Figures 1 and 2 show two snapshots of parts the documents used in the IMT evaluation.

The results of this experiment is show in Table 1.

The evaluation shows that none of the human raters rated the output to be *not understandable*. Importantly, the result shows that a significant percentage of the human raters have rated the output as *understandable*. However, 8% of the participants in the test, ranked the results as *completely understandable*.

Understandability	**Sorani to Kurmanji**	**Kurmanji to Sorani**
Not Understandable	0%	0%
Slightly Understandable	29%	21%
Understandable	63%	71%
Completely Understandable	8%	8%

Table 1: Understandability of the IMT output - The table shows that 82% of the human raters, rated the output of IMT to be quite understandable.

To the best of our knowledge, there is no baseline or golden-standard available for inter-dialect translations at the time of writing this paper. To illustrate, in the work by Zhang (1998) on inter-dialect MT, neither a quantitative evaluation of the developed system, nor any measures and baselines with which the system’s performance could have been evaluated, have been provided. This is, perhaps, because this work seems to be the first study of the kind. As an another example, Nakov and Tiedemann (2012) who have studied MT among closely-related languages have used BLEU (Papineni et al., 2002) as an evaluation method and compared their suggested approaches using the mentioned method. However, we were not able to apply BLEU for the reasons we mentioned in Section 3.2. Also in a recent work Shah and Boitet (2015) have used raw machine translation for translating Hindi tweets into English and have used the measure of *understandability* without referring to a certain baseline. Therefore we based our evaluation on the definitions that we suggested in Section 3.2.

5 Discussion

The experiment showed that the system performs at an acceptable level as about 82% of human raters rated the results as understandable. The evaluation also shows that none of the human raters rated the output to be *not understandable*. However, this rate for Sorani to Kurmanji is less than the rate for Kurmanji to Sorani outputs. The reason for this, as participants in the evaluation also confirm it, is because in the translation process, as the consequence of lack of underlying language resources, we could not apply the techniques of reordering the words and word alignment to make the output to completely conform with the Kurmanji structure. This causes the out-

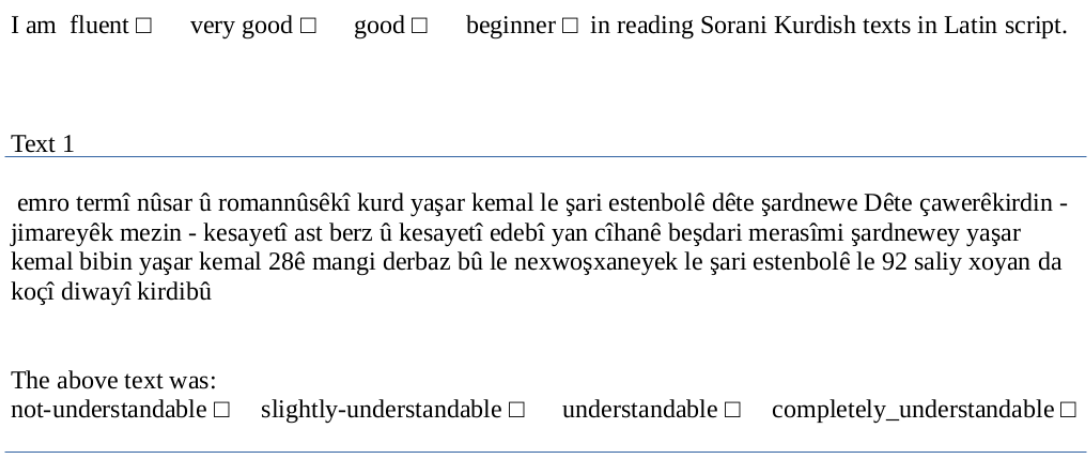

Fig. 1: IMT Evaluation - Kurmanji to Sorani

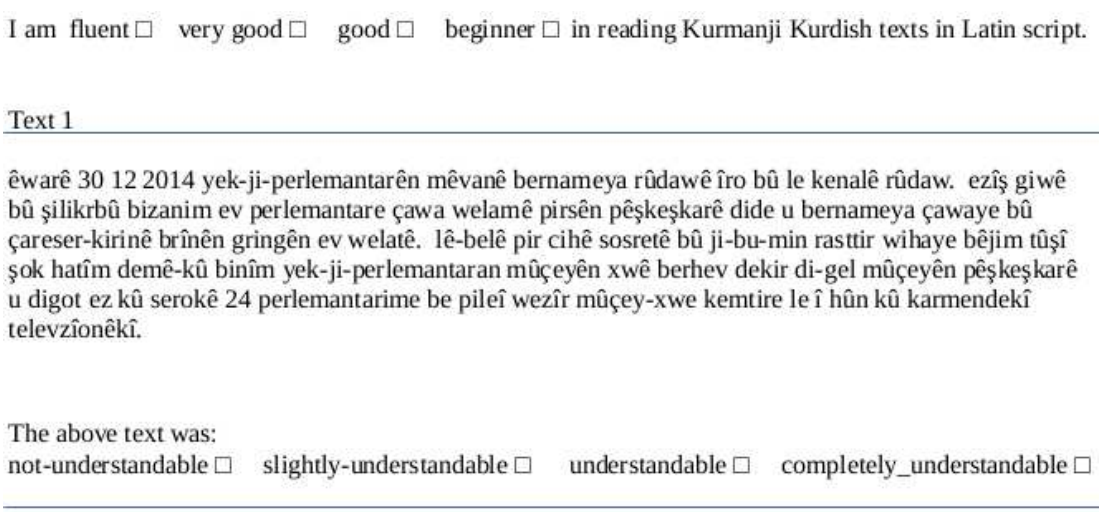

Fig. 2: IMT Evaluation - Sorani to Kurmanji

put to be seemed as an artificial and influent text which makes the evaluation rate different for Sorani to Kurmanji translation. But, because Sorani does not recognize genders and also its structure is more flexible as a result of borrowing more structures from other dialects, particularly, in the Iraqi Kurdistan region, where the evaluation has been conducted, the translated texts into Sorani have received better ratings.

The results showed that the method performs at an applicable level. However, we are also interested in finding the justifications for this fact from computational perspective. As Table 2 shows, the two bidialectal dictionaries do not share a high percentage common vocabulary. That is, perhaps the common vocabulary is not the only reason that justifies the acceptable performance of the system. Therefore, we will look into the Levenshtein distance between our bidialectal dictionaries to assess whether there is any correlation between this parameter and the our hypothesis about the efficiency of IMT.

Count	**Total**	**Kurmanji**	**Sorani**
Words	6792	2632	4160
Common Words	208	208	208
Percentage	3%	7%	5%

Table 2: The table shows the number of words attributed to each dialect alongside the common words among the dialects. It also shows the percentage of the common words to all words and total words in each dialect dictionary.

5.1 Levenshtein Distance and Intelligibility

We studied the results of our experiments from the perspective of Levenshtein distance in both bidalectal dictionaries in order to find any correlations between the efficiency of the suggested method and the similarities among the dialects vocabulary. Researchers in the NLP and CL have addressed the issue of intelligibility and similarity among languages and dialects from different points of view. For instance, Casad (1992) states that "the set of variables that underlie linguistic similarity are largely distinct from those that underlie intelligibility". Unlike "linguistic similarity", word similarity and word synonymy is one of the most computationally developed ideas. It has applications in several areas of NLP and CL such as Information Retrieval (IR), summarization, and MT. Two words are considered as similar if they share common meaning elements (Jurafsky and Martin, 2008). Word similarity has been investigated in different contexts. As a related example to this research subject, Bondi Johannessen et al. (2005) have investigated word similarity in the Scandinavian languages which are assumed as mutually intelligible dialects. Also Ljubšić and Kranjčić (2015) have studied Language Identification (LI) on Twitter using word similarity (Ljubšić and Kranjčić, 2015).

Word distance is another perspective of word similarity. From this perspective, one can measure the differences between two words (the "distance") instead of their similarities. That is, the less the distance between the words, the more similar the words are and the more the distance between te words, the less similar the words are. The Levenshtein distance measures the distance between the two words by counting the number of deletions, substitutions, and insertions that transfers one sequence into the other (Jurafsky and Martin, 2008). In the context of our experiment, Beijering et al. (2008) have applied this method in their study about the dialectal and inter-language intelligibility.

We calculated the Levenshtein distance among our bi-dialectal dictionaries entries. Figures 3 and 4 show the results of this calculation for Kurmanji-Sorani and Sorani-Kurmanji dictionaries respectively.

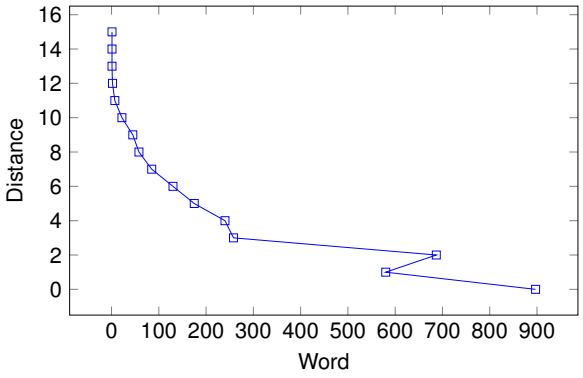

Fig. 3: Levenshtein Distance, Kurmanji-Sorani Bidialectal Dictionary - The plot shows the numbers of words with a certain distance in the dictionary.

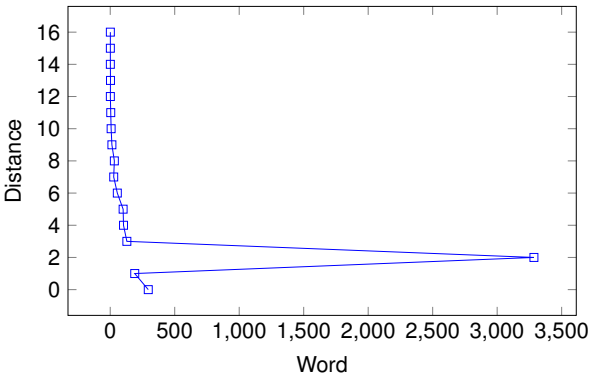

Fig. 4: Levenshtein Distance, Sorani-Kurmanji Bidialectal Dictionary - The plot shows the numbers of words with a certain distance in the dictionary.

As the Figures 3 and 4 show, except in a few cases, the plots indicate that majority of vocabulary of the two dialects have no more than 2 to 3 distances with each other. Based on this figures and the data of in Table 2 we suggest two reasons for the competency of our method. First, the Kurmanji and Sorani dialects are sharing a common vocabulary that although does not form a large portion of their lexicon, plays an important role as the basis for their lexicon structure. Second, the differences in a majority of the vocabulary that is used in social conversations differ by one or two letters, or sometimes just phonemes, for example, "şt̂ek" and "şêwezmanî" in Sorani versus "tşt̂ek" and "şêwezimanî" in Kurmanji. Further investigation is required perhaps with the help of Kurdish linguists to come up with solid conclusion in the future studies.

6 Conclusion

We implemented an Intralingual Machine Translation for translating texts in Kurmani to Sorani. We used word-for-word translation (literal or direct translation) among the dialects. The results were tested using human raters method. The experiments, according to different human raters, showed that this naive approach could provide a significantly intelligible results according to different human raters. The experiment also showed that this approach might be able to be considered as an immediate remedy for the current lack of corpus issue. In fact, the approach, if incorporated to the online resources, allow the speakers of one dialect to have access to the resources of the others with a reasonable degree of understandability. It also allows Kurdish computational linguists to focus on other aspects of the computational dialectology through studying the intelligibility issues.

Nevertheless, this study has not intended to consider the aesthetic aspects of the translation as an art, rather its intention is to evaluate the adequacy of this approach in the absence of large data that could support a high quality translation. Moreover, the outcome of this study might help other linguistics studies about the relation of the diverse dialects of a language such as Kurdish. The experiments showed that translated texts are understandable according to readers in 71% cases for Kurmanji and 79% for Sorani. They are slightly understandable in 29% cases for Kurmanji and 21% for Sorani.

However, there are several areas that we believe might be of interest as future work. For example, to investigate the extent to which the word-for-word translation of Kurdish dialect could be beneficial. For instance, to develop a system based on a shallow-transfer and rule-based approach using Apertium platform (Peradin et al., 2014) and to compare this method with the previous one in terms of the quality of the output, the speed of the system and the simplicity of reapplying the method to other Kurdish dialects. In addition, the evaluation approach can be changed from human rating to automated methods for example, BLEU (Papineni et al., 2002). Also to assess whether these approaches eliminates the role of parallel corpus in **intralingual** translation by adding grammatic rules, can be conducted as another research. Moreover, to analyze the case of word ambiguity in the implemented IMT, more investigation on the role of the Levenshtein distance, the implication of phonological/morphological differences, and similarities between the dialects are other areas that need to be studied further. Finally, the implementation of the method for translation between other Kurdish dialects could reveal more enlightening facts about the mutual intelligibility among these dialects. This also helps in better understanding the role of IMT with respect to making these dialects comprehensible among different audience, particularly web users.

Acknowledgments

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 who have provided constructive suggestions that have improved the final version of this paper.

References

Arafat Ahsan, Prasanth Kolachina, Sudheer Kolachina, Dipti Misra Sharma, and Rajeev Sangal. 2010. Coupling Statistical Machine Translation with Rule-based Transfer and Generation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas*.

Association for Computional Linguistics. 2015. *Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*. Association for Computational Linguistics (ACL), September.

Lisa Ballesteros and Bruce Croft. 1996. Dictionary Methods for Cross-lingual Information Retrieval. In *Database and Expert Systems Applications*, pages 791–801. Springer.

Karin Beijering, Charlotte Gooskens, and Wilbert Heeringa. 2008. Predicting intelligibility and perceived linguistic distances by means of the Levenshtein algorithm. *Linguistics in the Netherlands*, 15:13–24.

Janne Bondi Johannessen, Kristin Hagen, Åsne Haaland, Andra Björk Jónsdottir, Anders Nøklestad, Dimitris Kokkinakis, Paul Meurer, Eckhard Bick, and Dorte Haltrup. 2005. Named entity recognition for the mainland Scandinavian languages. *Literary and Linguistic Computing*, 20(1):91–102.

Eugene H. Casad. 1992. *Windows on bilingualism*. Summer Institute of Linguistics.

Sin-wai Chan. 2014. *Routledge Encyclopedia of Translation Technology*. Routledge.

Chin-Chuan Cheng. 1997. Measuring relationship among dialects: DOC and related resources. *Computational Linguistics & Chinese Language Processing*, 2(1):41–72.

Umîd Demîrhan. 2007. *Ferhenga Destî -kurdî bi kurdî - çapa duyem a berfireh (Kurdish to Kurdish Dictionary, expanded 2nd ed)*. weşanên SEWADê (*SEWAD Publication)*. In Kurdish.

Ali Farghaly and Khaled Shaalan. 2009. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):14.

Rebecca Fiederer and Sharon O´Brien. 2009. Quality and Machine Translation: A Realistic Objective. *Journal of Specialised Translation*, 11:52–74.

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

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

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

Mehrdad R. Izady. 1992. *The Kurds: A concise handbook*. Taylor & Francis.

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

Angela Kluge. 2006. RTT retelling method: An Alternative Approach to Intelligibility Testing. *Occasional paper. Dallas: SIL International*.

Kurdish Academy of Languages. 2016. The Kurdish Population.

Nikola Ljubšić and Denis Kranjčić. 2015. Discriminating between closely related languages on twitter. *Informatica*, 39(1):1.

Shervin Malmasi. 2016. Subdialectal differences in sorani kurdish. In *Proceedings of the VarDial Workshop*, Osaka, Japan, December.

Ayishe Mohammed Ali. 2008. *Kurdish-Kurdish Dictionary*. Ministry of Culture-KRG. In Kurdish.

Jeremy Munday. 2009. *Introducing translation studies: Theories and applications*. Routledge.

Preslav Nakov and Jörg Tiedemann. 2012. Combining Word-level and Character-level Models for Machine Translation Between Closely-related Languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 301–305, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kurdish Academy of Language. 2016. KAL Featured Articles.

Ergin Öpengin and Geoffrey Haig. 2014. Regional variation in Kurmanji: A preliminary classification of dialects. *Kurdish Studies*, 2(2):143–176.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Hrvoje Peradin, Filip Petkovsky, and Francis M. Tyers. 2014. Shallow-transfer rule-based machine translation for the Western group of South Slavic languages. In *Proceedings of the 9th SaLTMiL Workshop on Free/open-Source Language Resources for the Machine Translation of Less-Resourced Languages*, pages 25–30.

Joe E. Pierce. 1952. Dialect distance testing in algonquian. *International Journal of American Linguistics*, 18(4):203–210.

Erdal Ronahî. 2015. Lîsteya soranî-kurmancî (Soranî-Kurmanji Dictionary). In Kurdish.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic Identification of Arabic Dialects in Social Media. In *Proceedings of the First International Workshop on Social Media Retrieval and Analysis*, pages 35–40. ACM.

Lenhart Schubert. 2015. Computational linguistics. *The Stanford Encyclopedia of Philosophy*. Spring 2015 Edition.

Ritesh Shah and Christian Boitet. 2015. Understandability of machine-translated Hindi tweets before and after post-editing: perspectives for a recommender system. In *Tweet Translation Workshop 2015*, pages 44–50. CEUR-WS.org.

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

Kyumars Sheykh Esmaili. 2012. Challenges in Kurdish Text Processing. *arXiv preprint arXiv:1212.0074*.

Cecilia Szeto. 2000. Testing intelligibility among Sinitic dialects. In *Proceedings of ALS2K, the 2000 Conference of the Australian Linguistic Society*. Citeseer.

Chaoju Tang and Vincent J. van Heuven. 2009. Mutual intelligibility of Chinese dialects experimentally tested. *Lingua*, 119:24.

Chaoju Tang, Vincent J. van Heuven, et al. 2008. Mutual intelligibility of Chinese dialects tested functionally. *Linguistics in the Netherlands*, 25(1):145–156.

Carl F. Voegelin and Zellig S. Harris. 1951. Methods for determining intelligibility among dialects of natural languages. *Proceedings of the American Philosophical Society*, 95(3):322–329.

Wîkîferheng and Ferhenga azad (*Azad Dictionary*). 2015. Pêvek:soranî-kurmancî (*Amendment: Sorani-Kurmanji*). In Kurdish.

Joseph K. Yamagiwa. 1967. On dialect intelligibility in Japan. *Anthropological Linguistics*, pages 1–17.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann, editors. 2014. *The 1 st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, Dublin, Ireland, August. Association for Computational Linguistics (ACL).

Xiaoheng Zhang. 1998. Dialect mt: a case study between cantonese and mandarin. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 1460–1464. Association for Computational Linguistics.