Towards Machine Translation for the Kurdish Language

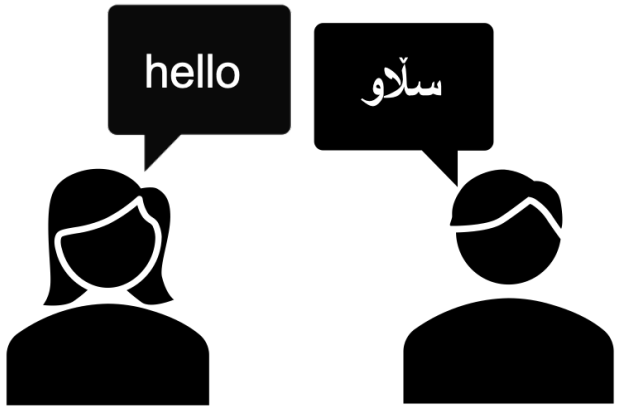
Sina Ahmadi, Mariam Masoud

Insight Centre for Data Analytics
National University of Ireland Galway

The 3rd Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2020) - AACL-IJCNLP

Table of Contents

Introduction



- machine translation (MT) is one of the major and most important sub-fields in natural language processing (NLP)
- languages are not equally addressed and challenging in MT
  - less-resourced languages: lack of parallel resources
  - language-specific features: morphologically-rich languages
  - availability of basic NLP tools, particularly tokenizers

Table of Contents

Objectives

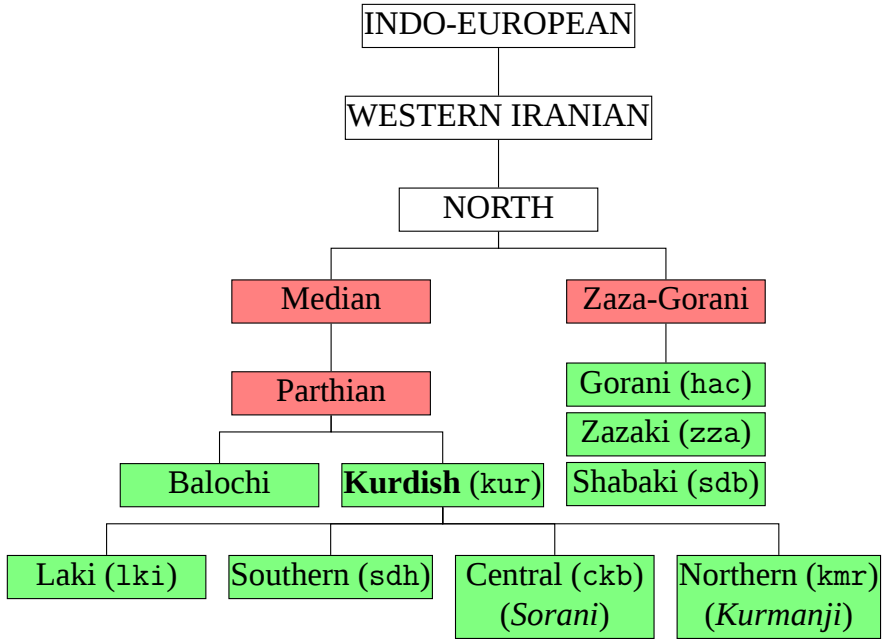
1. provide a description of the Sorani dialect of Kurdish and some of its linguistic features

2. shed light upon the MT task for Sorani by:
   - creating a basic setup for essential language processing tasks for MT in Kurdish
   - building and evaluating neural machine translation (NMT) systems for Sorani Kurdish

Table of Contents

Kurdish Language

- an Indo-European language
- spoken by 20-30 million speakers
- spoken in many dialects and subdialects (*dialects* or *languages*?)
- written in many scripts, among which the Latin-based and Arabic-based ones are still widely in use
- a less-resourced language



* Source: https://www.britannica.com/topic/Kurd

Kurdish in the MT realm

- very few previous studies addressing Kurdish MT
- In 2016, a rule-based MT system for Kurmanji and Sorani added to the Apertium project
- A general-purpose online service called `inKurdish`[1] using dictionary-based methods for translation
- In 2016, the translation service of Google, i.e. Google Translate, added Kurmanji Kurdish
- Kurdish language translation has been also of interest to many humanitarian organizations
- Shortly after our project in August 2020, the Microsoft Translation service added Sorani and Kurmanji[2]
- **still a long way ahead!**

[1] `https://inkurdish.com/`

[2] `https://www.bing.com/translator`

Sorani Kurdish

- mostly written in the Arabic-based script of Kurdish
- vastly spoken in the Kurdish regions of Iraq and Iran
- Some linguistic features:
  - a system of tense-aspect-modality and person marking (no grammatical gender)
  - a split-ergative language
  - a complex morphology with various clitics and affixes appearing in erratic patterns, particularly endoclitics

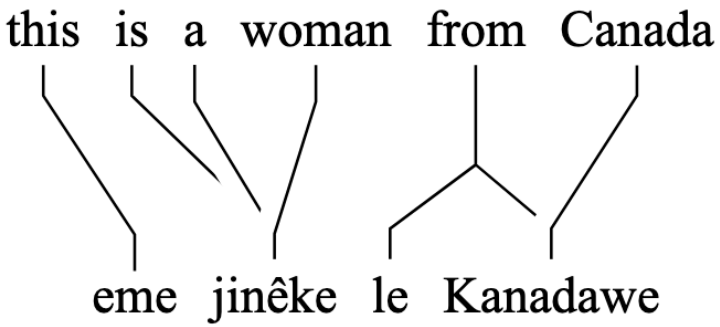
(1) *gułekanman hênan.* /. گوڵەکانمان هێنان
gułekanman hênan .
guł=ek-an=man hêna-in
flower.DEF.PL.1PL *bring*.PST.TR.ERG.3SG
'we brought the flowers.'

(2) *hênamanin.* /. هێنامانن
hênamanin .
hêna=man-in
bring.PST.TR.ERG.1PL.3SG
'we brought them.'

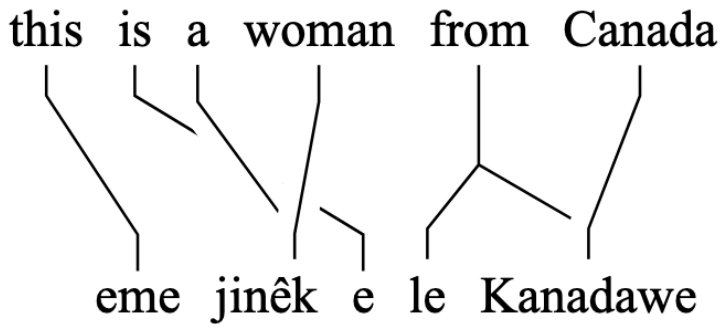
(3) *deçine małman.* /. دەچنە ماڵمان
deçin e małman.
de-çi-in=e mał=man.
go-PRS.PROG.3PL=to house.N.1PL.
'(they) go/are going to our house.'

Alignment in Sorani Kurdish

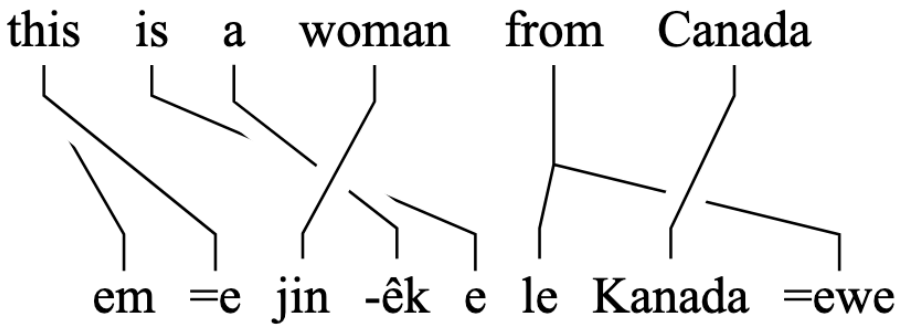
Kurdish has a subject-object-verb word order



(a) alignment at word level



(b) alignment at token level



(c) alignment at morpheme level

Available resources for Kurdish

- There has been an increasing number of resources created for the Kurdish language, such as
  - dictionaries (Ahmadi et al.,2019)
  - domain-specific corpora (Abdulrahmanet al.,2019)
  - folkloric corpus (Ahmadi et al.,2020a)
  - KurdNet–the Kurdish WordNet (Aliabadi et al.,2014)
- However, parallel corpora are more scarcely available:
  - Bianet (Ataman,2018): a parallel news corpus containing 6,486 English-Kurmanji Kurdish and 7,390 Turkish-Kurmanji Kurdish sentences
  - Ahmadi et al.(2020b) present a parallel corpus containing Sorani, Kurmanji and English parallel sentences
  - GNOME and Ubuntu localization files
  - Tanzil corpus provides translation of Qoranic verses in Sorani

Table of Contents

Data Preparation: datasets

Three parallel corpora:

- **Tanzil Corpus**
  - a collection of Quran translations compiled by the Tanzil project[3]
  - one translation in Sorani aligned with 11 translations in English
  - 92,354 parallel sentences with 3.15M words in the Sorani side and 2.36M words in the English side
  - religious context, noisy text, inconsistencies in writing Kurdish ”لوط” (Lot) vs ”لوت”
- **TED Corpus**[4]: 2358 parallel sentences in Sorani and English in a wider range of topics in comparison to Tanzil
- **KurdNet–the Kurdish wordNet** containing 4,663 definitions which are directly translated from the Princeton WordNet (version 3.0)

[3] http://tanzil.net

[4] https://wit3.fbk.eu

Data Preparation: datasets

In some cases, the alignments do not essentially correspond to a *sentence*:

Corpus	Language	tokens per line	characters per line
Tanzil	Kurdish	25.82	159.36
	English	27.96	134.72
Ted	Kurdish	69.21	441.93
	English	93.54	452.88
KurdNet	Kurdish	7.51	44.27
	English	8.51	49.14

Data Preparation: preprocessing

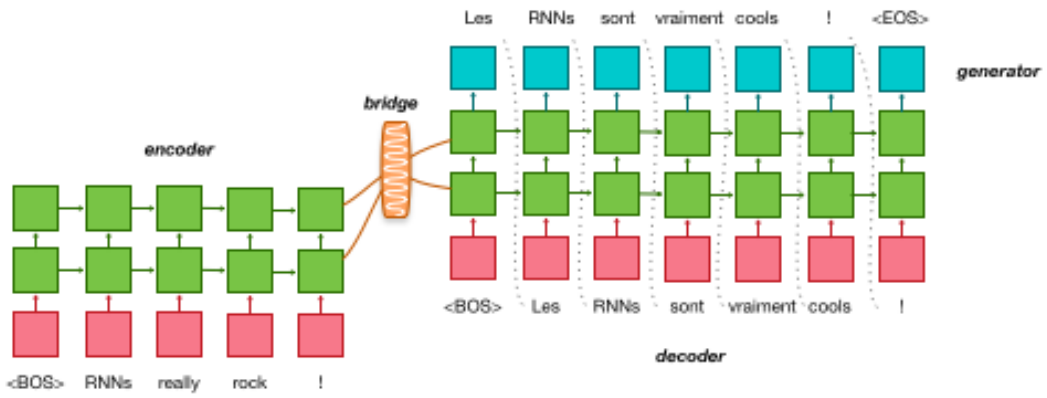
- remove non-relevant characters and clean the data
- remove additional interpretations between parentheses provided by the translator, e.g. “peace be upon him”
- unify the encoding of the characters by converting similar graphemes to unique ones, e.g. ”ك” and ”ي” respectively to ”ک” and ”ی”
- orthographic normalization, such as replacing initial ”ر” (r) with ”ڕ” (ř).
- removal of text within parentheses and truecasing (English data only)

Tokenization

- the task of tokenization is of high importance in various tasks in NLP, particularly in MT
- as of August, no tokenization tool was available for Kurdish
- created tokenization models using unsupervised methods: `WordPiece`, `Unigram`, `byte-pair-encoding (BPE)`, `WordPunct`

Tokenizer	Dataset	Number of tokens (sentences)				
		All	Train	Validation	Test 1	Test 2
BPE	Tanzil	3,335,725 (92325)	2,406,706 (66476)	296,874 (8308)	297,237 (8309)	334,908 (9232)
	TED	253,777 (2355)	185,258 (1697)	22,324 (212)	21,879 (211)	24,316 (235)
	KurdNet	46,384 (4659)	33,542 (3357)	4,054 (418)	4,178 (419)	4,610 (465)
	All		2,625,506 (71532)	323,252 (8940)	323,294 (8941)	363,834 (9932)
Unigram	Tanzil	3,365,517	2,428,059	299,634	299,765	338,059
	TED	260,015	189,879	22,860	22,377	24,899
	KurdNet	46,336	33,491	4,055	4,171	4,619
	All		2,651,429	326,549	326,313	367,577
WordPiece	Tanzil	3,348,264	2,415,538	298,174	298,321	336,231
	TED	247,865	180,822	21,773	21,615	23,655
	KurdNet	46,228	33,391	4,063	4,168	4,606
	All		2,629,751	324,010	324,104	363,742
WordPunct	Tanzil	2,909,512	2,098,910	258,926	259,381	292,295
	TED	250,617	183,596	21,886	21,353	23,782
	KurdNet	38,950	28,130	3,450	3,514	3,856
	All		2,310,636	284,262	284,248	319,933

NMT models



*

- py-Torch version of OpenNMT (Klein et al.,2017)
- deployed two variations of sequence to sequence NMT models:
  - Model 1: two LSTM (Long Short Term Memory) layers with 200 hidden units for both the encoder and the decoder
  - Model 2: the default OpenNMT model
- used the FastText pre-trained word vectors for Kurdish (Mikolov et al.,2018) and GloVe word embeddings trained on 6B tokens for English

Evaluation

- the performance of the models is evaluated using the following three evaluation metrics:
  - BLEU (Papineni et al.,2002) that matches n-grams
  - METEOR (Lavie and Agarwal,2007) based on the harmonic mean of precision and recall
  - TER (Snover et al.,2006): the cost of editing the output of the MT systems to match the reference
- create various datasets based on the tokenization techniques
- as the Tanzil corpus is remarkably larger, we create two separate test sets for the evaluation:
  - Test 1: we first set 10% of each dataset individually aside
  - Test 2: the remaining 90% of the data are then merged and shuffled and then split into 80%, 10% and 10% fro train, validation and test sets

Table of Contents

Results and Analysis

Corpus		Tokenization	Model 1						Model 2					
			ckb-en			en-ckb			ckb-en			en-ckb		
			BLEU	METEOR	TER	BLEU	METEOR	TER	BLEU	METEOR	TER	BLEU	METEOR	TER
Test 2	Tanzil	WordPiece	21.02	0.2400	0.61	51.44	0.3635	0.32	19.48	0.2310	0.64	50.48	0.3616	0.32
		Unigram	20.71	0.2381	0.60	50.21	0.3582	0.32	19.53	0.2320	0.63	50.95	0.3613	0.32
		WordPunct	**22.03**	**0.2454**	**0.58**	58.36	0.4120	**0.27**	20.42	0.2384	0.61	**59.28**	**0.4156**	**0.27**
		BPE	21.03	0.2392	0.61	50.04	0.3588	0.32	19.49	0.2315	0.63	50.28	0.3580	0.33
	KurdNet	WordPiece	5.86	0.1245	0.93	**3.90**	0.0910	**0.99**	**6.47**	**0.1297**	0.90	3.25	0.0904	1.01
		Unigram	5.88	0.1216	0.91	3.38	0.0884	1.00	6.15	0.1269	**0.89**	3.82	**0.0923**	1.00
		WordPunct	5.81	0.1169	0.90	2.57	0.0820	1.00	5.16	0.1242	0.90	2.82	0.0867	1.00
		BPE	6.32	0.1209	0.92	3.50	0.0853	1.00	6.39	0.1330	0.90	3.05	0.0826	0.99
	TED	WordPiece	**1.00**	**0.0875**	0.90	0.00	0.0378	0.99	0.74	0.0758	0.90	**0.05**	0.0383	0.99
		Unigram	0.88	0.0775	0.91	0.00	**0.0415**	**0.97**	0.89	0.0863	**0.89**	0.00	0.0397	0.99
		WordPunct	0.62	0.0720	**0.89**	0.00	0.0295	1.00	0.59	0.0712	0.90	0.00	0.0289	0.99
		BPE	0.92	0.0853	0.91	0.00	0.0353	0.99	0.75	0.0803	0.90	0.00	0.0298	0.99
Test 1		WordPiece	19.05	0.2242	0.65	46.47	0.3322	0.37	17.49	0.2153	0.67	45.23	0.3280	0.38
		Unigram	18.95	0.2235	0.63	45.24	0.3275	0.38	17.47	0.2156	0.66	45.83	0.3299	0.38
		WordPunct	**19.95**	**0.2276**	**0.61**	52.21	0.3726	**0.33**	18.50	0.2222	0.65	**52.94**	**0.3753**	**0.33**
		BPE	19.06	0.2233	0.63	45.13	0.3282	0.38	17.51	0.2157	0.66	45.14	0.3269	0.38

Results and Analysis

Quantitative Analysis

- in both Kurdish to English and English to Kurdish translations, the WordPunct performs better
- Surprisingly, Model 2 which is trained with more hyper-parameters, performs better only in English
- in Test 2, all the setups fail to translate KurdNet and TED corpora
  - imbalance of the data
  - type of sentences in KurdNet and the quality of alignments in TED
  - domain-specific terms used in the Kurd-Net and TED corpora

Qualitative Analysis

- system translations often carry meaning in a comprehensible way
- the trained models capture information regarding synonyms or semantically-related words, e.g. ‘knowledge’ is translated as زانست (*zanist*) ‘science’ vs زانیاری (*zanyarî*) ‘knowledge’

Results and Analysis

Input (Tanzil)	when they came in to him , and said , salam ! he answered ; salam , and said : you are a people unknown to me .
Reference	کاتێک کتوپر خۆیان کرد بە ماڵدا و وتیان : سڵاو ، ئەویش وتی ، سڵاو لەئێوەش بێت ، هەرچەندەناتانناسم .
System translation	کاتێک چوون بۆ سەردانی و وتیان : سڵاو ، ئەویش وتی ، سڵاو لەئێوەش بێت ، هەرچەندەناتانناسم .
Back-translation	when (they) went to visit him/her and said : hi , then (he) said, hi to you too, although(I)donotknowyou .
Input (TED)	all the knowledge and values shared by a society
Reference	تەواوی زانست و بەهایانەی کە کۆمەڵ تێدا هاوبەشن
System translation	هەر زانیاری و ئامێرێکی تەواو بوون .
Back-translation	all the knowledge and a tool of finishing .
Input (KurdNet)	a structure consisting of a room or set of rooms comprising a single level of a multilevel building
Reference	پێکهاتەیەک کە لە ژووڕێک یان چەند ژوور دروست بووە و لە سەر نهۆمێکی بینایێکی چەند نهۆمیە
System translation	پێکهاتەیەک کە لە زنجیرەیەک یان لە دیوارێ هاتۆتە دەرێ کە لە ئاستێکی گەورەترداید
Back-translation	a structure that has come out from a chain/a set or a wall that is within a bigger level

Table of Contents

Conclusion and Future Work

Conclusion:

- MT systems for Sorani Kurdish translation are presented
- various aspects of building an MT system for Kurdish are highlighted, including the lack of basic language processing tools
- Two limitations of the current work:
  - a base-line system
  - further experiments with respect to hyper-parameters

Future work:

- morpheme-based translation (Luong et al.,2019)
- usage of lexicons for compensating the scarcity of resources for Kurdish (Zhang and Zong,2016)
- monolingual sequence-to-sequence pre-training techniques, such as MAsked Sequence toSequence pre-training (MASS) (Song et al.,2019)

Thanks for your attention!

Our datasets and models are available at:
`https://github.com/sinaahmadi/KurdishMT`