A Tokenization System for the Kurdish Language

Sina Ahmadi

Insight Centre for Data Analytics, National University of Ireland Galway, Ireland

Objectives

- Carry out a preliminary study on the task of tokenization for the **Kurdish language**
- Describe the Kurdish language and **word boundary** in it
- **Create a tokenization system** for two of the Kurdish dialects, i.e. **Kurmanji** and **Sorani** which are respectively written in a **Latin** and an **Arabic-based** script
- Compare the performance of our lexicon-based approach with unsupervised tokenization methods

Introduction

A text, as the input of text processing applications, is composed of **a string of characters** and is interpreted based on the way it is segmented. **Words** and **sentences** are two segments in a text which carry meaning at different levels. Although the **boundaries of words** and sentences are specified to some extent in some scripts, e.g. by using whitespaces and punctuation marks, finding such boundaries is **a non-trivial task**.

Given the recent advances in NLP and artificial intelligence, tokenization is considered a solved problem and has been efficiently addressed for many languages. Although methodologies and approaches in tokenization of one language might be applicable to and beneficial for another language, **linguistic and orthographic issues** can make tokenization a **language-specific problem**.

Kurdish and its Word Boundary

- Kurdish is a **less-resourced Indo-European language** spoken by 20-30 million speakers in the Kurdish regions of Iran, Iraq, Turkey and Syria [1]
- Northern Kurdish (Kurmanji), Central Kurdish (Sorani), Southern Kurdish and Laki are the dialects of Kurdish
- An Arabic-based alphabet is widely used for Sorani and Southern Kurdish while a Latin-based is used for Kurmanji
- There is no consensus regarding what is meant by a **standard writing system** or **orthography**
- In both the Latin-based and Arabic-based scripts of Kurdish, whitespaces are used for delimiting word boundaries. However, none of these delimiters are deterministic for word boundary in Kurdish due to:
  - **Orthographic Inconsistencies**: various variations are found with respect to writing a specific word in Kurdish texts
  - **Excessive Concatenation**: many short tokens, such as adpositions and copula may merge with other word forms without proper spacing
  - **Compound Words**: Having a relatively few number of around 300 single-word verbs, i.e. verbal lexemes, Kurdish extensively uses compound forms to develop its vocabulary. Finding boundary of compound forms is a non-trivial task as well

Approach

As a preliminary study, we focus on the application of a lexicon of lemmata and morphological analysis for tokenization of Kurdish texts. Moreover, we follow the common practices in tokenization, such as **detecting digits, dates, URLs** and **punctuation marks** as distinct tokens. This sub-task is called "normalization prior to tokenization" [2].

```
{
    "bi-can-û-bên": {
        "token_forms": [
            "bicanûbên",
            "bi canûbên",
            "bican ûbên",
            "bi can ûbên",
            "bicanû bên",
            "bi canû bên",
            "bican û bên",
            "bi can û bên"
]}}
```

Listing 1: A Kurmanji compound lemma and its possible forms in the lexicon in JSON

Lexicon To develop a lexicon for our task, we use the lexicographic material of Free-Dicts (`https://freedict.org`) and the Kurdish Wiktionary, Wîkîferheng (`https://ku.wiktionary.org`). Overall, **8,180** and **9,970** headwords are collected in Sorani and Kurmanji among which **1,513** and **1,507** lemmata are compound forms. We follow these steps to create our lexicons:

- Cleaning and normalization the characters
- Transliterate scripts
- Retrieve headwords consisted of more than one word and follow a standard convention by separating all compound forms by a hyphen (-)
- For each compound form, we create all the possible forms **with and without a space**

Morphological Analyzer We create a morphological analyzer to create simpler word forms by striping concatenated morphemes.

Experiments

- **Data annotation**: Manually annotate 100 sentences in Sorani and Kurmanji in the Text Corpus Format
- **Tokenization models**: We create our baseline model using the `WordPunct` tokenizer of NLTK + four unsupervised neural models:
  - `WordPiece`
  - **Byte Pair Encoding** (`BPE`), **unigram language model** (`Unigram`) and `Word` model

```
{
  "ئاخر-و-تۆخ": {
        "token_forms": [
                "ئاخرو تۆخ",
                "ئاخروتۆخ",
                "ئاخر و تۆخ",
                "ئاخر وتۆخ",
                "ئاخروو تۆخ",
                "ئاخروتۆخ"
                ]
}
}
```

Listing 2: A Sorani compound lemma and its possible forms in the lexicon in JSON

Evaluation

Due to the limited advances in Kurdish language processing, we evaluate our tokenization **as a component alone** and not in an end-to-end setup. We evaluate all the models using BLEU-n (from 1 to 4) and accuracy. **We demonstrate that our system outperforms the other methods with a remarkable difference in the accuracy**.

Dialect	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Acc.
Sorani	0.98	0.95	0.91	0.87	30.44
Kurmanji	0.97	0.94	0.91	0.87	31.38

The following is a demo of the system detecting compound forms and separating them correctly:

- Sorani
  - Raw: "دوا کەوتنی شیوازەکانی بەرهەمهێنان"
  - Tokenized: ['__دوا-کەوتن__ی_', '_شیواز_ەکان_ی_', '__بەرهەم-هێنان__']
- Kurmanji
  - Raw: "endamên encûmena wezîrên herêma Kurdistanê"
  - Tokenized:['_endam_ên_', 'encûmen_a', '_wezîr_ên_', '_herêm_a_', '_Kurdistan_ê_']

References

[1] Sina Ahmadi, Hossein Hassani, and John P. McCrae. Towards Electronic Lexicography for the Kurdish Language. In *Proceedings of the eLex 2019 conference*, pages 881–906, Sintra, Portugal, 1–3 October 2019. Brno: Lexical Computing CZ, s.r.o.

[2] Rebecca Dridan and Stephan Oepen. Tokenization: Returning to a long solved problem—a survey, contrastive experiment, recommendations, and toolkit—. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–382, 2012.

Use the tool

- Annotated resource: `https://github.com/sinaahmadi/KurdishTokenization`
- Tool: `https://github.com/sinaahmadi/klpt`

Insight SFI Research Centre for Data Analytics · Data Science Institute, Institiúid Eolaíochta Sonraí · OÉ Gaillimh NUI Galway