Enhanced Urdu Word Segmentation using Conditional Random Fields and Morphological Context Features

Aamir Farhan, Mashrukh Islam and Dipti Misra Sharma
Language Technologies Research Centre
International Institute of Information Technology
Hyderabad, India
{aamir.farhan, mashrukh.islam}@research.iiit.ac.in, dipti@iiit.ac.in

Abstract

Word segmentation is a fundamental task for most of the NLP applications. Urdu adopts Nastalique writing style which does not have a concept of space. Furthermore, the inherent non-joining attributes of certain characters in Urdu create spaces within a word while writing in digital format. Thus, Urdu not only has space omission but also space insertion issues which make the word segmentation task challenging. In this paper, we improve upon the results of Zia, Raza and Athar (2018) by using a manually annotated corpus of 19,651 sentences along with morphological context features. Using the Conditional Random Field sequence modeler, our model achieves F_1 score of 0.98 for word boundary identification and 0.92 for sub-word boundary identification tasks. The results demonstrated in this paper outperform the state-of-the-art methods.

1 Introduction

Word segmentation is the first and foremost task for NLP applications, such as sentence parsing, part-of-speech tagging and machine translation. Word segmentation can be explicitly challenging for languages which do not have a delimiter to mark word boundary in their writing system. Urdu is one such language which is written in Arabic script using Nastalique writing system. Conventionally, Nastalique writing system adopts writing in a continuous fashion without any space characters. However, when Urdu is written in digital format, white space is used to mark word boundary as well as sub-word boundary in order to get correct shaping of a non-joining character, as discussed in the next section. Due to this absence of an objective delimiter, Urdu faces complex issues in word segmentation. This paper explains the problem of word segmentation in Urdu and presents an enhanced model for solving the tokenization problem using Conditional Random Fields (CRFs) along with morphological context features.

2 Literature Survey

Several rule-based methods have been used for tokenization issues in Urdu. Durrani and Hussain (2010) broadly descirbed the Word Segmentation issues and proposed various hybrid methods to solve the problem. One of them being a rule-based dictionary lookup with maximum matching. Another hybrid method involved statistical modeling using n-grams along with hueristics to identify the top 10 segmentations of an input string of Urdu characters. Their hybrid method achieved error detection rate of 85.8% for space omission and space insertion errors. Zia et al. (2018) proposed a CRF based model for Urdu Word Segmentation along with a publicly available corpus.

3 Urdu Writing System

Urdu is written in cursive Arabic script, also known as Nastalique. Urdu characters aquire different shapes according to the their respective positions in a word. The characters can be divided into two categories, *joiners* and *non-joiners.* Depending upon their position in a sequence, a joiner can have four shape variants: 1) initial 2) medial 3) final and 4) isolated form. Non- joiners only have two forms, final and isolated.

Conventionally, Urdu does not have a concept of white space as a word delimiter. A white space is instead used while typing to prevent a character from joining to its subsequent character. This ensures that correct shape of a character is maintained. For example, Urdu typists learn to insert a white space within the word خوش قسمت (Fortunate) to get the correct shape of ش. Without space, it appears like خشکسمت which is visually incorrect.

Thus Urdu writing system has space insertion issues when a white space is inserted within a word to get correct shape of a character and space omission issues when the joiners connect to each other in the Nastalique writing style.

4 Word Segmentation Model

To solve the aforementioned issues in the Urdu writing system, we present the word segmentation problem as a sequence labelling task where each character in the input sequence is assigned one of the following labels: 1) B : Beginning of a word 2) S : Beginning of a sub-word 3) O : Others

4.1 Corpus

The system presented by Zia et al. (2018) was trained using a manually crafted corpus of 4,325 sentences which is relatively small when compared to benchmark corpora of languages like Arabic and Chinese. We manually annotated a much bigger corpus of 19,651 sentences from Urdu news journals. We used white space to mark word boundary and Zero Width Non-Joiner (ZWNJ) to mark sub-word boundary, consistent with the rules proposed by Rehman et al. (2011). The corpus covered most Urdu morphological constructions. For training and testing purposes, we have split the data in the following terms : 17,400 sentences for training and 2250 for testing.

4.2 Features

We crafted context features in terms of N-grams to capture the morphological context of a character in a sequence. The features involve unicode class of the character along with N-grams up to four preceding and four succeeding characters. We also incorporated a boolean feature in terms of whether a character belongs to Urdu script or not. For example, for the character م in the phrase الزام لگاکر will have the following set upto 4-grams as part of its features

['c=م', 'c-ا', 'c-ام', 'c-زام', 'c-لزام', 'c-الزام', 'c+ل', 'c+مل', 'c+م لگ', 'c+م لگا', 'c+م لگاک']

4.3 Model

Since the word segmentation problem is now transformed to a sequence labeling problem, we adopted a Conditional Random Field (CRF) model proposed by Lafferty et al. (2001). CRF is a framework for building probabilistic models to label sequence data. The model is defined as $P(\mathbf{y_1...y_n}|\mathbf{X})$ where **X** is a sequence of input and $\{y_1...y_n\}$ is a sequence of predicted labels. The labels, as described above, belong to the set, L = {B, S, O}.

5 Results

For evaluation metrics, we used precision, recall and F_1 measures. On a testing data set of 2250 sentences, our model achieved F_1 score of 0.98 for label B (word boundary), 0.92 for label S (sub-word boundary) and 0.99 for label O (others). The results show significant improvement in the F_1 scores as compared to the previous methods primarily because of two factors:

- Significantly larger data size and richness of the new training corpus.
- The new engineered features which consider N-grams up to four preceding and four succeeding characters.

Label	Precision	Recall	F_1 Score
O	0.99	0.99	0.99
B	0.98	0.98	0.98
S	0.90	0.93	0.92

Table1: Test results corresponding to each label

	O	B	S
O	137294	1126	3
B	1080	47586	29
S	5	16	300

Table2: Confusion matrix for sequence labeling

6 Limitations and Future Scope

The model proposed in this project is trained using a manually crafted corpus which is relatively small compared to segmentation benchcmark corpora of resource rich languages like Arabic and Chinese. The model also heavily depends on manually engineered features for determinig word and sub-word boundaries. We have explored the usage of neural models like Bi-LSTM for this task. Since the neural models are data hungry, we plan to expand our annotated corpus and explore suitable character embeddings which can be fed to a Bi-LSTM RNN. We also plan to tag special grammatical constructions in Urdu like the *Izafa* constructions and add this label to our annotation schema as well.

References

Durrani, N., & Hussain, S. (2010, June). Urdu word segmentation. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 528-536). Association for Computational Linguistics.

Zia, Raza and Athar (2018). Urdu Word Segmentation using Conditional Random Fields(CRFs). Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics.

Afzal, M., & Hussain, S. (2001). Urdu computing standards: development of Urdu Zabta Takhti (UZT) 1.01. InMulti Topic Conference, 2001. IEEE INMIC 2001. Technology for the 21st Century. Proceedings. IEEE International(pp. 216-222). IEEE.

Cai, D., & Zhao, H. (2016). Neural word segmentation learning for Chinese. arXiv preprint arXiv:1606.04300.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Monroe, W., Green, S., & Manning, C. D. (2014). Word segmentation of informal Arabic with domain adaptation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Vol. 2, pp. 206-211).

Rehman, Z., Anwar, W., & Bajwa, U. I. (2011). Challenges in Urdu text tokenization and sentence boundary dis-ambiguation. In Proceedings of the 2nd Workshop on South Southeast Asian Natural LanguageProcessing (WSSANLP) (pp. 40-45).

Ping, G., & Yu-Hang, M. (1994). The adjacent matching algorithm of Chinese automatic word segmentation and its implementation in the QHFY Chinese-English system. In Proceedings of the 1994 International Conference on Chinese Computing, Singapore (Vol. 301, p. 94).

Wong, P. K., & Chan, C. (1996, August). Chinese word segmentation based on maximum matching andword binding force. In Proceedings of the 16th conference on Computational linguistics-Volume 1 (pp. 200-203). Association for Computational Linguistics.

Green, S., & DeNero, J. (2012, July). A class-based agreement model for generating accurately inflected transla-tions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Pa-pers-Volume 1 (pp. 146-155). Association for Computational Linguistics.

Sornlertlamvanich, V. 1995. Word Segmentation for Thai in a Machine Translation System (in Thai), Papers on Natural Language Processing, NECTEC, Thailand