Discriminating Similar Languages: Persian and Dari

Shervin Malmasi
Centre for Language Technology
Macquarie University, Sydney, Australia
shervin.malmasi@mq.edu.au

ABSTRACT

Although widely-studied in recent years, Language Identification (LID) systems for determining the language of input texts often fail to discriminate between similar languages like Croatian-Serbian and Malay-Indonesian. This has brought attention to the task of discriminating similar languages, varieties and dialects – including a recent shared task [3].

Persian (also known as Farsi) and Dari (Eastern Persian, spoken predominantly in Afghanistan) are two close variants that have not hitherto been investigated in LID and we report the first results on this pair. Dari is a low-resourced but important language, particularly for the U.S. due to its ongoing involvement in Afghanistan, which has led to increasing research interest [1].

We developed a corpus of 28k sentences (14k per-language) and using character and word *n*-grams, we discriminated them with 96% accuracy. Out-of-domain cross-corpus evaluation was conducted to test the discriminative models' generalizability, achieving 87% accuracy in classifying 79k sentences from the Uppsala Persian Corpus. Feature analysis revealed lexical, morphological and orthographic inter-language differences.

Further to determining document languages, LID has applications in character encoding detection, statistical machine translation, inducing dialect-to-dialect lexicons and authorship profiling in the forensic linguistics domain. In Information Retrieval it can help filter documents (e.g. news articles or search results) by dialect.

LID can also be used in other Natural Language Processing tasks, including the adaptation of tools like part-of-speech taggers for low-resourced languages [2]. Since Dari is too different to directly apply Persian resources, the distinguishing features identified through LID can assist in adapting existing resources.

BODY

Language Identification methods using surface features can distinguish close linguistic variants Persian and Dari with 96% accuracy.

REFERENCES

[1] S. Condon, L. Hernandez, D. Parvaz, M. S. Khan, and H. Jahed. Producing Data for Under-Resourced Languages: A Dari-English Parallel Corpus of Multi-Genre Text. 2012.

[2] A. Feldman, J. Hana, and C. Brew. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proceedings of LREC*, pages 549–554, 2006.

[3] M. Zampieri, L. Tan, N. Ljubešic, and J. Tiedemann. A report on the DSL shared task 2014. *COLING 2014*, page 58, 2014.

Volume 3 of Tiny Transactions on Computer Science
This content is released under the Creative Commons Attribution-NonCommercial ShareAlike License. Permission to make digital or hard copies of all or part of this work is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
CC BY-NC-SA 3.0: http://creativecommons.org/licenses/by-nc-sa/3.0/.