Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_r_00521

Book Review

Automatic Language Identification in Texts

Tommi Jauhiainen, Marcos Zampieri, Timothy Baldwin, and Krister Lindén
University of Helsinki, George Mason University, MBZUAI, and University of Helsinki

Springer (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 44), 2024, xiv+148 pp;
ISBN 978-3-031-45821-7;
ebook, ISBN 978-3-031-45822-4;
doi:10.1007/978-3-031-45822-4

Reviewed by
Tom Lippincott
Johns Hopkins University

Language identification (LI) for text data, in the ideal scenario, determines the human languages employed at every location in a corpus. In practice this often means choosing the likeliest language at the document level: this is already quite useful, e.g when presenting a webpage to the user and deciding a) whether to translate it and b) which model to use for that purpose. However, nuances like code-switching (language alternation), dialect variation, and ambiguously-short content are increasingly common with the ubiquity of digital communication like text messaging and micro-blogs. Geographic areas like Africa and the Indian subcontinent bring enormous linguistic diversity and flexibility that breaks the document-level LI paradigm. While standard references (Jurafsky and Martin 2023) introduce LI, touch on these subtleties, and often present related methods and models in other contexts, *Automatic Language Identification in Texts* is specifically dedicated to LI in its full practical variety.

In the course of producing a broad and thorough survey, perhaps the most striking takeaway from Jauhiainen et al. is the chaotic state of research on this critical task. This might be due to the view that, for digitally well-attested languages occurring in domains with monolingual documents of at least modest length, LI is solved: these circumstances are common, and the emphasis on massive data sets can make the rarer cases seem less important. When challenges arise in specific, applied downstream research, they are often addressed in an ad hoc fashion, such as through active learning techniques for gathering human annotations or linear programming to incorporate prior knowledge (Lippincott and Van Durme 2016), without consolidation into broader outcomes for the research community. Throughout *Automatic Language Identification in Texts*, the authors have the consistent goal of improving this situation. The book is structured into six chapters:

Chapter 1 introduces the history of LI, stretching from early feature-engineering approaches to still-standard models based on character n-grams closely related to fundamental models of communication (Shannon 1948), and the burgeoning collection of shared tasks aimed at specific domains, such as ancient scripts or regional dialects. Unlike much of machine learning for natural language processing tasks, traditional models have remained highly competitive for LI compared to deep neural networks: perhaps data sparsity prevents effective training, or traditional features are already well-suited for LI. Downstream use-cases and challenges are summarized, with copious citations to prior and ongoing work.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_r_00521

Chapter 2 begins with the authors' efforts to standardize the discourse around LI by specifying a common notation that subsumes the variety employed in the literature. While the notation is a modest shift from those that treat data as a sequence of fully-distinct documents, treating documents as boundaries within a single large sequence of characters consolidates the spectrum of methods that will be covered. In terms of linguistic features, the focus is on character n-grams, and the authors address several standard concerns: weighting, smoothing, and incorporating linguistic knowledge. The latter is particularly interesting and perhaps under-explored, since there is often less practical motivation to move beyond the immediate use-case and consider e.g. the phylogenetic structure of world languages. The bulk of the chapter is devoted to describing a wide range of classification methods that employ these features, some standard (e.g. logistic regression, naive Bayes), others the specific ensembles or statistical tests adopted by existing research.

Chapter 3 addresses evaluation, the other end of the experimental pipeline that requires standardization. While a handful of metrics have been used historically, most research has converged on macro balanced F-score, which equally weights precision and recall as well as performance on each language. In the absence of a clearly-articulated reason, this is the most even-handed approach. The bulk of the chapter is devoted to a survey of standard data sets and shared tasks, both historical and ongoing. This is a useful reference for researchers in search of venues aimed at their specific goals, or looking for broader patterns in outcomes.

Chapter 4 considers the primary axes that may elevate LI from "solved" to "challenging": language similarity, low-resource languages, orthographic systems and variation, short text, and code-switching. Some of these involve questions of representation: what do we treat as a "language"? What is the "correct" label of a short text that's valid in multiple languages, such as "quando?", which is a common question in Portuguese and Italian? How should one label a text containing multiple languages, such as "I'll ask mi hombre next time I see him"? Chapter 5 then considers the pursuit of a maximally-general model capable of characterizing massive collections of heterogeneous content, unknown languages, and domain shift.

Chapter 6 discusses several prominent or otherwise compelling uses of LI, from the pragmatic needs of machine translation to subtle tasks like determining the native language based on characteristic patterns in a second language. For instance, corpora of writing from known L2 speakers of English are widespread due to the popularity of English as a second language throughout education, allowing the study of orthographic mistakes grounded in phonetic properties of a native language. Stylistics and authorship attribution share useful features with LI, as they strive to avoid learning *topical* properties that are often correlated with language.

The authors conclude by reiterating the diversity of phenomena that existing LI techniques rarely treat as first-order challenges (until they become immediately relevant), and the difficulty of drawing broader conclusions from the current literature. The book effectively catalogues these challenges and heterogeneity while also providing a stable reference for the community working to organize and extend research in this area. This is useful for several audiences and purposes:

- Students seeking to understand the history and landscape of LI
- Researchers hoping to unify or extend existing methods
- Practitioners or stakeholders who need to select and justify an approach to a specific task

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_r_00521

The only notable "limitation" of the book is in fact endemic to the topic: the poorly-mapped variety of LI research is naturally going to show through any thorough survey. The authors are up front about this state of affairs and succeed at improving on it.

References

Jurafsky, D. and J.H. Martin. 2023. Speech and Language Processing.
Lippincott, Tom and Benjamin Van Durme. 2016. Fluency detection on communication networks. pages 1025–1029.
Shannon, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Tom Lippincott is Associate Research Professor in Computer Science and Director of the Center for Digital Humanities at Johns Hopkins University. He has published on language and dialect identification, particularly under conditions of extreme data sparsity, and explored approaches that combine standard techniques with augmentations from linear programming, signal processing, and active learning. Lippincott's email address is tom.lippincott@jhu.edu.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_r_00521