

Language Identification

Timothy Baldwin, Baden Hughes,
Jeremy Nicholson



THE UNIVERSITY OF
MELBOURNE

Language Identification (LangID): Introduction

- Naive definition: determine the language l_i document D is written in
simple classification problem
- Considered largely “solved” problem
- N.B. written vs. spoken language identification very different tasks

Current Reality of LangID

- Focus on “high(er)-density” languages
emphasis on **token** rather than **type** accuracy
- Focus on web documents
- Most methods based on character/byte n -grams

Basic Questions

- How much data do we need to be able to do accurate LangID?
- What is the intrinsic difficulty of the LangID, and do we really need “top-end” machine learning solutions?
- How much data is “enough” to be able to perform accurate LangID?
- Can we perform LangID equally with word lists (work types) and document statistics (work tokens)?

Open Issues

Supporting Minority Languages

- How well do existing techniques support language identification for languages which form the bulk of the more than 7000 languages identified in the Ethnologue?

Open Class Language Identification

- Can we treat LangID as an open-class classification problem?

$$\arg \max_{c \in C} lm(c, D) \text{ vs. } \arg \max_{c \in C \cup C'} lm(c, D)$$

Sparse or Impoverished Training Data

- What is the performance of the variety of LangID systems in environments where the amount of gold standard data for training is small (e.g. 50/100/250 words or 50/100/250 characters)?

Multilingual Documents

- Can we move away from a one-to-one view of LangID to a one-to-many view?
 - ★ finer granularity (e.g. sentence, paragraph, section)
 - ★ in quantitative terms (e.g. a document is 3% French, 95% English and 2% Italian)

Standard Evaluation Corpora

- Can we come up with a standard evaluation corpus which is:
 - ★ multilingual
 - ★ representative of linguistic diversity
 - ★ representative of “interesting” text sources/language resources

Performance Evaluation Criteria

- Can we move away from IR-style evaluation criteria to produce something more representative of reality?
 - ★ gradated judgements for source language
 - ★ gradated judgements for resource type
 - ★ possibly micro-level markup of the location of different languages in the document

Effects of Preprocessing

- Is it possible to talk about LangID independent of preprocessing?
 - ★ stemming
 - ★ stop word removal
 - ★ case folding

Non-Roman Script / Multi-script

- Is all that glitters really Roman script based? Can we move away from the Latin script-centric view of LangID?
 - ★ Languages with multiple scripts
 - ★ (Non-Latin) scripts with multiple languages

Legacy and Non-Standard Encodings

- How do we deal with legacy/non-standard encodings?
 - ★ is it equally difficult to identify a language across multiple encodings?
 - ★ are all encodings equally (dis)similar?
 - ★ encoding vs. linguistic similarity

Exploiting the Linguistic Content of Documents

- Can a richer model of document “semantics” help leverage more reliable LangID?

Legacy Data vs. Lack of Support

- How do we deal with legacy (incl. ill-formed) web documents, ad hoc character encodings, and orthographies without encodings?

Conclusion

- There is still lots of fun to be had in the LangID space, and still significant progress to be made
- Collaborative resource development/sharing (the construal of LangID as a “shared task”) will help bring LangID into the mainstream