Language Identification in Bilingual Texts for Linguistic Data Extraction

Terrence Szymanski
tdszyman@umich.edu

This talk presents a novel language identification task: token-level language ID in bilingual documents, specifically in the context of a system for extracting linguistic data from digital documents. The overall extraction system operates by first identifying instances of foreign-language text in the document, then identifying nearby reference-language text that is a translation the foreign text. The goal, illustrated in Figure 1, is to produce paired texts and translations as output, forming a parallel corpus that can then be used for linguistic research. Whereas parallel corpora currently only exist for a few dozen languages, the present work has the potential to create similar corpora (though smaller in size) for a much larger number of languages.

Previous work on language ID has generally focused on monolingual texts, using a sample of text from each different language to train a classifier (Kruengrkai 2006). More related work (Xia et al., 2009) has addressed the question of how to identify the language of previously-identified segments of text within multilingual documents. The approach in this paper differs in that language identification is done at the word level, not the document level, simultaneously segmenting the bilingual text into smaller monolingual spans. Furthermore, in our approach we assume that one of the languages in the bilingual document is a known reference language (e.g. English), but we do not assume any prior knowledge of the second, target language. Two different approaches are implemented and compared to human inter-annotator agreement rates as well a dictionary-based baseline. The first approach is fully unsupervised, using a probabilistic model of English along with a threshold cutoff value to distinguish English tokens from non-English. The second approach uses a small amount of hand-labeled data to train a SVM classifier for the target language.

The output from the language ID system is used as input for a second stage of processing in which adjacent spans of English text are considered as candidate translations for each span of foreign-language text. All spans of two or more foreign-identified words are considered, and the two immediately adjacent spans of text (both preceding and following) are proposed as candidate translations. These are used to train a statistical translation model of the foreign language, and all of the candidate translations are scored according to this model. The higher scoring candidate is chosen as the predicted translation. Examples of the predictions are given in Figure 2. Non-adjacent translations (example 1) and errors in identifying the full foreign text span (example 3) are representative of the challenges faced by the current system.

Manual evaluation of a 100-sentence subset of predicted parallel text instances from a grammar of Santhali (Skrefsrud 1873) showed that 99% of the predicted foreign texts were correctly identified as Santhali, and 28% of the predicted English translations were correct. The most common error was that the system cuts off the foreign spans too early, meaning that despite its high precision, the language ID component could stand to be improved in terms of recall (estimating recall is difficult in the absence of a labeled evaluation data set). There are a number of possible routes towards improving accuracy: using commercial software to improve OCR quality, incorporating typological and layout information into the classier, and tuning the parameters of the statistical models. However, the current system is more than sufficient for the task of data discovery, since a user can always verify the data against the original source document.

This line of work is inspired by the Online Database of Interlinear Text (ODIN) (Lewis & Xia 2010), and made possible by the availability of scanned books in library digitization projects. By applying natural language processing methods to digital documents containing linguistic data, projects like this one can create richer resources and facilitate linguistic analysis by providing linguists a centralized and searchable interface to linguistic data from a variety of sources. Existing resources such as The Ethnologue (Lewis 2009), and WALS (Dryer & Haspelmath 2011) provide valuable metadata about languages, but not actual linguistic utterances. Furthermore, in order to make progress in the area of data-intensive experimental linguistics (Abney 2011), it is important to have access to machine-readable data sets from as many languages as possible.

Presently, the system is being developed and tested on a small corpus of seven grammars. However, it has the potential to scale to include any scanned document on the web, and could be extended to collect data from websites and other forms of text. The current focus is on out-of-copyright texts, and the ultimate objective of this work is to produce a corpus that can be shared online with the research community.

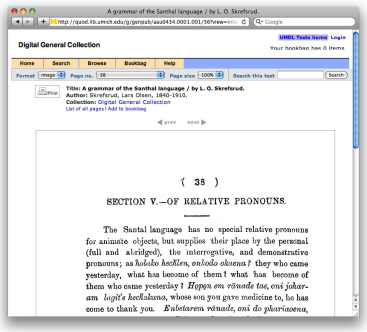

Figure 1: : Example of input (scanned document on the left) and expected output (parallel text on the right) of the extraction process.

1.
```
had struck him.     had struck him.     he had struck hitn.
DUAL.               DUAL.                 DUAL.
I D-al-a1,kat'-ti;4-ta- Dal-akat'-li.-tcth'- Paset'-e-dat-a~cat'-liti..
lt-1can-a-e,  He    kan-A-han-e,   If   tcth~1oan,   Perhaps
had   struck  us    he had struck us    he   had  struck us
```

2.
```
strike.
INCHOATIVE PAST.
Dal-Jko-dagido11-kan-tahVkan,
Tlhey whom they were about
to strike.
OPTATIVE.
```

3.
```
oni hola-m del-led-e, what has become of him    whom   you
saw yesterday? This is much more elegant and certainly more
correct than to say: oni hola-m diel-ed-e-a, oni do okare,
for the latter means literally: you saw him yesterday, what
has become of him?
```

Figure 2: Three examples of predictions made by the system (foreign text in bold, translation underlined).

References

S. Abney. Data-intensive experimental linguistics. *Linguistic Issues in Language Technology*, 6. 2011.

C. Kruengkrai, P. Srichaivattana, V. Sornlertlamvanich, and H. Isahara. Language identification based on string kernels. In KICSS. 2006.

W. D. Lewis and F. Xia. Developing ODIN: A multilingual repository of annotated language data for hundreds of the world's languages. *Literary and Linguistic Computing*. 2010.

F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics.*, 29(1):19–51. 2003.

L. O. Skrefsrud. *A grammar of the Santhal language*. Benares: Medical Hall Press. 1873.

F. Xia, W. Lewis, and H. Poon. Language id in the context of harvesting language data off the web. In *EACL*:870–878. 2009.

M. S. Dryer and H. Martin (eds.) 2011. The World Atlas of Language Structures Online. Munich: Max Planck Digital Library. 2011.

M. P. Lewis (ed.) Ethnologue: Languages of the World. Online version. `http://www.ethnologue.com/` Dallas, TX: SIL International, sixteenth edition. 2009.