Towards unsupervised extraction of linguistic typological features from language descriptions

Søren Wichmann
Leiden University
wichmannsoeren@gmail.com

Taraka Rama
University of Oslo
tarakaramadaiict@gmail.com

1 Introduction

Manual encoding of typological databases is a tiresome procedure that takes large amounts of time. Bender (2016) reviews recent efforts in extracting typological features from interlinear glossed text (Lewis and Xia, 2010), Bible corpora (Östling, 2015; Malaviya et al., 2017), and sources such as morphologically annotated resources and treebanks (Bjerva and Augenstein, 2018).

However, there is a lack of publications describing the application of NLP techniques to extract typological features directly from language descriptions contained in grammar books, dissertations, and linguistics articles. Collections of such descriptive sources are accumulating as PDFs (including many from scans) that have subsequently been OCR'ed. In this paper, we describe our first attempt at building an NLP pipeline that extracts typological features from OCR'ed linguistic descriptions.

2 General approach

Our approach to extracting features in WALS (Dryer and Haspelmath, 2013) from the OCR'ed texts consists of two steps. First we detect that part of a text which is most likely to contain a description of a given WALS feature. Next, we try to solve the classification problem consisting in extracting exactly one feature value from the target text chunk. That is, unseen chunks hypothesized to discuss a given WALS feature are matched with the general patterns associated with a specific feature value found in the training set. We use a training set of $10,000$ feature - value - source combinations.

3 Identifying text chunks containing feature descriptions

Initial preprocessing included cleaning the texts for noisy content. The relevant online WALS chapters were parsed and each put into a text file. Five different off-the-shelf keyword extraction methods were run on the WALS chapters. Their outputs are, respectively

1. POS-tags, yielding nouns and their frequencies
2. Collocations and co-occurrences and their frequencies
3. Keywords and their ranks using the Textrank algorithm
4. Keywords and their frequencies using rapid automatic keyword extraction (RAKE)
5. Noun and verb phrases and their frequencies

All our experiments involving keywords were performed using the R binding for the UDPipe package (Straka and Straková, 2017).[1]

For each combination of feature and language description the five above-mentioned keyword-extraction methods were applied to successive windows of 5 chunks of the description in order to find the combination of keyword and vector method most adequate for identifying a text chunk as discussing a particular WALS feature. For each window and keyword method, the distance was measured to the WALS chapters using 8 different standard vector distances (or similarities converted to distances): Chebyshev, correlation, cosine, Euclidean, Jaccard, Jensen-Shannon, Manhattan, and Soergel, in addition to a new one, called pJaccard.

[1] https://www.r-bloggers.com/an-overview-of-keyword-extraction-techniques/

For each combination of 5 keyword-extraction methods and 9 vector distances the text chunk with the smallest distance to the WALS chapter—the target chunk—was found.

For 345 of the sources a page number reference was given at least once. Ideally, one would like to use this information to see if the target chunk was found on the indicated page and use that as a criterion for evaluating the method of finding the target chunk. But splitting the 345 relevant OCR'ed descriptive sources into pages is not easily done.

The shape that descriptive sources take once we have converted them to R objects is as lines of text chunks carrying indices corresponding to each chunk, and what a line is depends on the nature of the OCR'ed text—typically it is a paragraph if the text was born digitally or one line of a printed text if the text came from a scan. The only resemblance these indices bear to page numbers is that the order of the indices and the order of the page numbers are perfectly correlated. We took advantage of this property and checked whether the indices corresponding to lines were ordered in the same way as page numbers when several WALS features came from one and the same source. The Spearman Rank correlation, ρ, was obtained for each case where one descriptive work was the source of $N(> 2)$ WALS feature values for a given language, and an average of the ρ values weighted by N was used as a yardstick for the performance of the different combinations of keyword extraction methods and vector distances for identifying relevant target chunks.

After running this experiment on around 10% of the training set, it became clear that the best performing combination of keyword extraction methods was the POS-tag method used with either Jensen-Shannon or cosine. While Jensen-Shannon performed a bit better, both were in the same ballpark, with ρ close to 0.1.

4 Identifying feature values in target text chunks

Having identified the text chunks that are potential sources for WALS features, the next step is to identify the WALS feature *value* directly from the text. The original training set was reduced to only those text chunks having a smaller than average Jensen-Shannon divergence to their WALS chapters. This served to get rid of all non-English texts and some more texts that can be assumed to be noisy for other reasons.

We divided the original (reduced) training set into a training and development set, where, for each combination of feature and value found, one half of the instances were assigned to the training set and the other half to the development set. When a combination occurred only once it was excluded. When a combination occurred an uneven number of times the training set was made to be one item bigger than the test set. Subsequently we compared the text chunks for each member of the development set (target chunks) with each text in the test set describing the relevant WALS feature (identifier chunks). The value of the best matching identifier chunk was identified as being the value of the target chunk. For this comparison we first extracted noun keywords and then applied both the cosine and Jensen-Shannon distances. The latter turned out to lead to more correct assignments.

5 Evaluation procedure

The baseline for the evaluation was defined as a situation where a random assignment is correct in proportion to the frequency of the given value in the test set. Taking the WALS feature 9A 'The Velar Nasal' as an example, let the values 'initial velar nasal', 'no initial velar nasal', and 'no velar nasal' occur respectively 6, 2, and 10 times in the training set. Then a random choice of, say, the value 'initial velar nasal' would count as 6/18 = 0.33 correct predictions in the baseline. The number of true baseline predictions for all development set members was summed up and compared to the sum of actual, true predictions.

6 Results and prospects

The baseline for correct feature value assignments is 44.09%. Using the cosine distance to identify feature values gave 44.40% correct answers, which is marginally above the baseline. Using Jensen-Shannon divergence (J-S) gave 45.72% correct answers, which is a clear improvement over the cosine. Moreover, an improvement for J-S could be obtained by requiring the distance between the target chunk and the best identifier chunk to be small. Thus, we get 45.9% correct assignments requiring J-S <0.5, and the amount of correct assignments increases monotonically as J-S decreases, up to 61.11% correct assignments for J-S <0.1.

There are many improvements to be made to

the various parts of the pipeline. We envisage that given such improvements we may be able to guess a feature value correctly close to 2/3 of the times when the data conditions are adequate, but we probably cannot expect anything better than that. Given the completely unsupervised approach and the early stage of the research this nevertheless seems worthwhile reporting. Future research will be aimed at improving the pipeline and combining it with other approaches.

7 General implications

The highly general and unsupervised approach taken in this paper is chosen with a view to wider applications to the extraction of information from technical literature where general descriptions (e.g., medicinal handbooks, codes of law), cases (e.g., patient records, legal proceedings), and categorical values (e.g., diagnoses, verdicts) need to be matched. Since there is nothing specific to linguistics in our procedure we hope that it may be of potential relevance to other fields of science and the humanities where texts and their interpretation play a central role in the development of the discipline.

References

Emily M Bender. 2016. Linguistic typology in natural language processing. *Linguistic Typology*, 20(3):645–660.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916.

Matthew Dryer and Martin Haspelmath. 2013. The world atlas of language structures online. Leipzig. Max Planck Institute for Evolutionary Anthropology.

William D Lewis and Fei Xia. 2010. Developing odin: A multilingual repository of annotated language data for hundreds of the world's languages. *Literary and Linguistic Computing*, 25(3):303–319.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535.

Robert Östling. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 205–211.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.