From Language Descriptions to Language Facts

Anonymous authors

Abstract

The diversity of the world’s 7 000 languages embodies a wealth of information on the communication machinery inside our heads as well as the history of populations. Traditionally language comparison has been done manually by humans reading grammatical descriptions, but the number of languages and books is now far beyond human capacities. In the present paper we report on experiments to transform a multilingual corpus of over 10 000 raw-text grammatical description into a collection of directly comparable “facts” for each language. The outcome database of facts is meant to serve for human investigation and human-designed machine comparisons. Therefore it is important that the collection of facts is human-interpretable and hence pure black-box solutions are not of interest. We report on successive experiments using Tf-Idf, Tf-Idf-inspired Term Extraction, Semantic Similarity and Multi-Lingual Semantic Similarity inspired techniques to build representations of the world’s languages. Since historical language relationships are broadly known, and, on the whole, language similarity co-varies with historical proximity, the extracted language facts can be evaluated in terms of how well they reflect known historical relations. As expected, accuracy improves with increased sophistication from simple word counts to tf-idf, semantically, and multilingually informed word counts, modestly but substantially advancing the frontier on this novel task.

1. Introduction

The diversity of the 7 000 languages of world represents an irreplaceable and abundant resource for understanding the unique communication system of our species. Rather than to study just one language, such as English, through the comparison of many languages we are better equipped to trace the history of the populations that speak them as well as to understand the processing machinery of our brains [Evans and Levinson, 2009].

For most of its history, comparison of languages has been carried out by humans qualitatively. This research framework is naturally bounded by the limits of human capacities, and comparison is now increasingly carried out with the aid of linguistic databases, notably the World Atlas of Language Structures (WALS, wals.info), the South American Indigenous
Manual curation of cross-linguistic grammatical data starts from grammatical descriptions. For example, each of the below works provide descriptive data on the grammar of one of the many minority languages of the world:


Human reading, although ideal in many ways, is very labour-intensive and potentially involves subjectivity. A database database treating a single linguistic topic for some 200 languages is typically the size of a PhD project, whereas the world has 7 000 languages and there is grammatical information for over 4 500 [Hammarström, 2021b]. Grammars are complex objects, and it remains to be seen how many thousands of features are needed to obtain an approximate characterization at any level of detail.

At the same time, a corpus of some 10 000 grammatical descriptions is now available in digital form for data mining research purposes [Virk et al., 2020a]. Extracting information from descriptive grammars is a relatively novel task and work so far has only sought to find answers to specific questions supplied from the outset [Virk et al., 2019, Wichmann and Rama, 2019, Macklin-Cordes et al., 2017, Virk et al., 2017, Hammarström et al., 2021]. In the present paper, we seek to automatically construct a comprehensive list of characteristics making up an the entire profile of a language [Virk et al., 2020b] in addition to populating it for the languages at hand.

In general terms, the problem addressed here can be formulated as follows. There is a set $D$ of raw-text descriptions of entities from a set $S$, such that each $d \in D$ mainly describes exactly one $s \in S$. If a term $k$ describing a property of objects in $S$ occurs in a document $d$ to a significant degree, the object $s$ described in $d$ actually has the property signalled by $k$. We seek to extract a profile $P$ with a suitable collection of properties that, in some manner, relate to properties signalled by such keywords. Further, given a document $d \in D$, we seek to populate a profile $p_s$ for the $s \in S$ described therein.

---

Since historical language relationships are broadly known, and, on the whole, language similarity co-varies with historical proximity, the extracted language profiles can be evaluated in terms of how well they reflect known historical relations. For example, English, German and Hindi belong to the Indo-European language family, but English and German are historically closer, in virtue of belonging specifically to the Germanic subfamily. Comparing the language profiles wholesale, English and German should then be closer than either is to Hindi. A special case of this principle is that different descriptions of one and the same language should be more similar to each other than to descriptions of other languages.

Different kinds of grammatical comparison yield different insights into the mystery of language and methods for comparison have to be masterminded by humans on data understandable to humans [Haspelmath, 2020]. We are therefore not interested in pure black-box solutions, e.g., neural networks, if they were to produce higher accuracy in some sense. Consequently, the present paper describes experiments exploiting relatively well-understood techniques such as Tf-Idf and Vector Space lexical semantics.

Languages can be compared both on the lexical and grammatical level. The grammars that constitute our raw data, by definition, contain predominantly grammatical information, but may also include smatterings of texts, lexicon, comparative, ethnographic and other information [Hammarström, 2021a]. We thus expect a successful extraction to be dominated by grammatical characteristics. Extraction of specifically lexical data is preferably done on dictionaries as the source data and is not targeted in the present paper [Segerer, 2016, Kamholz et al., 2014].

2. Previous Work

Given the novelty of the the task of automatically extracting information from descriptive grammars, only a few embryonic approaches have appeared [Virk et al., 2019, Wichmann and Rama, 2019, Macklin-Cordes et al., 2017, Virk et al., 2017, Hammarström et al., 2021]. All of them require the user to submit a word, phrase or description of the sought after feature but retrieve the result with varying depths of analysis of the source documents and with varying amounts of training data required. As noted above, the task addressed in the present paper also involves extracting the list of features itself and consequently only shallow and unsupervised methods can come into question for populating the feature list.

These premises of the task also apply to other domains and texts than linguistics, e.g., ethnographic descriptions. Judging from the surveys of Nasar et al. [2018] and Firoozeh et al. [2020], the premise that each \( d \in D \) mainly describes exactly one \( s \in S \) is not dominant across scientific domains. Consequently most work in information extraction from scholarly documents has focussed on the broader tasks of extracting key-insights and salient keywords from scientific documents. We are not aware of any work in other domains on the specific task addressed in this paper.

3. Data

The data for the experiments in this essay consists of a collection of over 10 000 raw text grammatical descriptions digitally available for computational processing [Virk et al., 2020a]. The collection consists of (1) out-of-copyright texts digitized by national libraries,
The collection has been OCRed using ABBYY Finereader 14 using the meta-language as recognition language. The original digital documents are of quality varying from barely legible typescript copies to high-quality scans and even born-digital documents. The OCR correctly recognizes most tokens of the set meta-language, but, particular to our collection, most documents contain a fraction of tokens which do not belong to the set meta-language but to the minority language(s) being discussed. These tokens are typically recognized poorly, as expected from the dictionary/training-heavy, contemporary techniques for OCR. We cannot easily improve on the OCR on a scale relevant for the present collection but some post-correction of OCR output very relevant for the genre of linguistics is possible and advisable (see Hammarström et al. [2017]). The bottom line is that searches for meta-language terms are relatively unimpaired by OCR but parsing of full sentences may falter.

Table 1: Meta-languages of the grammatical descriptions in the present collection.

<table>
<thead>
<tr>
<th>Meta-language</th>
<th># lgs</th>
<th># documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>3497</td>
<td>7284</td>
</tr>
<tr>
<td>French</td>
<td>826</td>
<td>1323</td>
</tr>
<tr>
<td>German</td>
<td>620</td>
<td>813</td>
</tr>
<tr>
<td>Spanish</td>
<td>394</td>
<td>808</td>
</tr>
<tr>
<td>Russian</td>
<td>288</td>
<td>498</td>
</tr>
<tr>
<td>Chinese</td>
<td>180</td>
<td>234</td>
</tr>
<tr>
<td>Portuguese</td>
<td>141</td>
<td>274</td>
</tr>
<tr>
<td>Indonesian</td>
<td>130</td>
<td>210</td>
</tr>
<tr>
<td>Dutch</td>
<td>113</td>
<td>171</td>
</tr>
<tr>
<td>Italian</td>
<td>92</td>
<td>141</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

archives, scientific societies and other similar entities, (2) texts posted online with a license to use for research, usually by university libraries and non-profit organizations (notably the Summer Institute of Linguistics), and (3) texts under publisher copyright where quotations of short extracts are legal. For each document, we know the language it is written in (the meta-language, usually English, French, German, Spanish, Russian or Mandarin Chinese, see Table 1), the language(s) described in it (the target language, typically one of the thousands of minority languages throughout the world) and the type of description (here restricted to grammatical descriptions only, but may include comparative studies, phonological descriptions, dictionaries etc more generally). The collection can be enumerated using the bibliographical- and metadata contained in the open-access bibliography of descriptive language data of Glottolog 4.4 (glottolog.org, Hammarström et al. 2021).
For Chinese, we employed the the Jieba tokenizer\textsuperscript{2} and normal punctuation-whitespace delimitation for the others. Words in all-capitals\textsuperscript{3} were kept as such while the remaining words are all lowercased.

4. Experiments

We present a series of experiments of increasing complexity to transform a raw text grammatical description $g$ of a language $L_g$ to a vector representation $X_g$. Vector representations of $g_1, g_2$ can be directly compared via Euclidean distance $d(X_{g_1}, X_{g_2})$.

Given the distance measure, each approach can be evaluated in terms of how well it reflects known historical relationships. Using Glottolog 4.4 (\texttt{glottolog.org}, Hammarström et al. 2021) as our database on historical relations, we sample 1000 triples from related languages featured in extraction experiments. Accuracy is then measured as the proportion of triples where $L_{g_1}$ and $L_{g_2}$ are historically more closely related to each other than to $L_{g_3} \Rightarrow d(X_{g_1}, X_{g_2}) < d(X_{g_1}, X_{g_3}) \land d(X_{g_2}, X_{g_3}) < d(X_{g_1}, X_{g_2})$. The random baseline is thus $\frac{1}{3}$. It is important that only triples of related languages are considered. Since the bulk of the extracted features may be structural in nature, it is possible that a language is more similar to an unrelated language than to a distant relative, and no penalties should be incurred if this is the case.

The methods are described and commented on below, and the accuracies are collected in Table 2.

**Frequency baseline:** As a baseline, we can represent each grammatical description as a raw count vector of the 10,000 globally most common terms. The accuracy of the frequency baseline hovers around the random baseline across all languages. In other words, the raw frequency counts of grammars provide little discriminatory power.

**Tf-Idf:** As a first step towards utilizing information asymmetries we consider Tf-Idf weighted vectors of sizes 1,000 and 10,000. The accuracy of Tf-Idf is also hovering close to the random baseline, with little consistent difference to the raw frequency approach.

**Doc2Vec:** Another possibility is to exploit distributional symmetries between words to create semantically informed document embeddings [Le and Mikolov, 2014]. We consider vector sizes 1,000 and 10,000. To save on time and memory, we discard words occurring less than 5 times but otherwise use the default settings in the \texttt{gensim} implementation [Řehůřek and Sojka, 2010]. The accuracy of Doc2Vec slightly improves on the previous approaches on the languages with the most data, but is largely similar.

**SIL-GOLD:** To restrict attention to words that are known to be specific to linguistic description we consider terms from the manually curated SIL Glossary of Linguistic Terms\textsuperscript{4} and the General Ontology for Linguistic Description (GOLD, Farrar and Langendoen 2003) totalling 2714 terms. We build vectors in two ways, one using the raw frequencies of occurrences of the SIL-GOLD terms, and one scaled by the logarithm

\textsuperscript{2} https://github.com/fxsjy/jieba
\textsuperscript{3} Modern linguistic descriptions contain so-called glossing, typically attained by capitalized abbreviations and these may be important to retain as such.
\textsuperscript{4} https://glossary.sil.org/ accessed 2021-05-12.
of the frequencies. The raw frequencies modus shows little improvement in accuracy (0.35) but the logarithm-scaled is a step up (0.40). Hence, restricting the view to linguistic descriptive terms reduces noise somewhat, but only if frequency extremities are levelled out by taking the logarithm. We only report the results for English, assuming the results would be similar for other languages if these lists were simply translated.

**Grammar features:** A list of terms relating to linguistic description can be extracted from the grammars themselves, competing with the manually curated lists. The heuristic is that a descriptive linguistic term should occur in on few pages across many grammars. E.g., descriptive terms such as classifier, laryngeal, inverse, deixis, etc. will likely be discussed in a specific section of many grammars. In contrast, (i) common general language words (e.g., the, a, in etc.) will occur on most pages of most grammars, and (ii) uncommon general language words (e.g., butcher, disruption, winston, etc.) occur on few pages in few grammars, and (iii) specific words relating to the description of a language (e.g., the language name, the name of a nearby river, specific forms of a specific language) occur on many pages of few grammars. Consequently, we can measure the “descriptivity” $D(t)$ of a term $t$ as the sum, across all grammars, of the inverse of the number of pages of occurrence of $t$.

$$D(t) = \sum_{g \in G} \frac{1}{\text{# pages in } g \text{ where } t \text{ occurs}}$$

For each meta-language, we consider the 10 000 most descriptive terms and retain the individual descriptivity value $D(t)$ for each term. A grammar is represented as the vector of each $D(t)$ multiplied by the logarithm of frequency of $t$. The automatically extracted list of grammar features replicate the results for the manually curated list (for English), showing that such external resources need not be relied on.

**Semantically-weighted grammar features:** Most of the 10 000 most descriptive terms are semantically related to various degrees, including terms that are near-synonymous. To conserve attention to independent features and discount multiple related features we arrange the list of descriptive terms in a tree based on their semantic similarity. First, semantic similarity for each language is obtained via Word2Vec [Mikolov et al., 2013b] as implemented in gensim [Rehúřek and Sojka, 2010] with 5 as the minimum count and otherwise default settings. Next, the terms on the grammar-features list are hierarchically clustered into a tree (Ward’s method). The descriptive terms are then re-weighted according proportionally according to the tree. Results show the semantic re-weighting to add another step of improvement over the naive features.

**Multilingual semantically-weighted grammar features:** All the vector embeddings so far have been specific to the (meta-)language of the grammar. Using the technique of Mikolov et al. [2013a] we may map the semantic vector spaces to the English one (with the most underlying data) and thence compare grammars written different languages freely. We otherwise retain the semantically-weighted grammar features method for embedding a grammar. While not unambiguous, the results indicate an
improvement when multilingualism is exploited. Small changes in the results may also be due to a richer set of multilingual grammar triples now available. Closer inspection is needed to assess the quality of semantic embeddings and translation quality across the languages.

Language descriptions are heterogeneous and grammars are complex objects to describe, so we should not expect there to be simple reductions. With increased sophistication towards mimicking independent grammatical features as commonly found in human-curated databases, accuracy does improve, albeit modestly. While clearly better than a frequency baseline, there is still a long way to go to perfectly replicate known historical relationships. In this setting, we could of course train a Neural Network embedding toward this goal to possibly obtain higher accuracy. However, our goal is a human-interpretable database obtained by automatic means, rather than prediction accuracy of opaque full-scale representations.

As expected from informal knowledge of descriptive materials, there are few discernable differences relating to the meta-language. The differences that do exist may also be due rather to dataset sizes and implicit regional biases than to meta-language itself.

5. Conclusion

There are good prospects for building cross-linguistic databases from the rich legacy of descriptive linguistic literature. The bulk of relevant documents are available in digital form for research purposes. We have shown that word co-occurrence based techniques can improve on the baseline of this novel problem and that the multilingual setting can be exploited, rather than combatted. Nevertheless, language descriptions are heterogeneous and grammars are complex objects to describe, so counting words in more or less advanced ways can only achieve so much. Hence, the present experiments should be seen as tools towards data collection in a less laborious way. The NLP techniques employed can be developed further and visualization/browsing tools are needed to oversee the large amount of data involved in the experiments.
Acknowledgments

Omitted for blind review.

References


