Unsupervised Language Filtering using the Latent Dirichlet Allocation

Wei Zhang[1], *Robert A. J. Clark*[2], *Yongyuan Wang*[1]

[1]Ocean University of China, Qingdao, China, 266100
[2]the CSTR, University of Edinburgh, United Kingdom, EH8 9AB

weizhang@ouc.edu.cn, robert@cstr.ed.ac.uk

Abstract

To automatically build from scratch the language processing component for a speech synthesis system in a new language a purified text corpora is needed where any words and phrases from other languages are clearly identified or excluded. When using found data and where there is no inherent linguistic knowledge of the language/languages contained in the data, identifying the pure data is a difficult problem.

We propose an unsupervised language identification approach based on Latent Dirichlet Allocation where we take the raw n-gram count as features without any smoothing, pruning or interpolation. The Latent Dirichlet Allocation topic model is reformulated for the language identification task and Collapsed Gibbs Sampling is used to train an unsupervised language identification model. We show that such a model is highly capable of identifying the primary language in a corpus and filtering out other languages present.

Index Terms: Language Filtering, Language Purification, Language Identification

1. Introduction

This paper concerns Language Identification in the context of 'purifying' a text corpus to determine which sentences are in the primary language of the corpus and contain no foreign words or phrases. This is a requirement for building the language processing front-end of a speech synthesis system entirely automatically in a new language where linguist resources other than the text are unavailable.

Language identification is usually viewed as a form of text categorization, Several kinds of classification approaches have used to identifying the language of documents: Markov Models combined with Bayesian classification [1], Discrete Hidden Markov Models [2], Kull-back Leibler divergence–namely relative entropy [3], minimum cross-entropy [4], decision trees [5], neural networks [6], support vector machines [7], multiple linear regression [8], centroid-based classifications [9] and improvements to the previous method [10]. Other work include conditional random fields [11] and minimum description length with dynamic programming [12].

These methods are all supervised and require clean editorially managed corpora for training. They are appropriate only for a limited number of languages, and require relatively large-sized documents. [13] demonstrate that "the task becomes increasingly difficult as we increase the number of languages, reduce the amount of training data and reduce the length of documents".

There have been some attempts to solve the problem of annotating training corpora. [14], in their multi-lingual speech synthesis, used the phonemes, words and sentences multilayer identification, and a combination of morphological and syntactic analysis. This kind of domain specific, sophisticated-design language identification is difficult to extend to the general situation. [15] according to the methodology of *Web As Corpus* [16], collect very large-scale multi-linguistic corpora and conduct their annotation, then train their *LangID.py* tool using domain adaptation, to provide an off-the-shelf tool for general language identification. The accuracy is still affected if the style of the documents to be identified is inconsistent with the training corpus. To address this issue, [17] studied language identification of eBay and twitters postings; he utilizes the initial and final words of postings and the corresponding site information for the initial annotation, and then bootstraps a supervised learning approach to achieve positive results. [18] with Twitter and Facebook posting language identification, also uses a bootstrap method where a trained supervised model, built from a Wikipedia corpus is tuned by fusing it with the Tweets location feature.

These approaches demonstrate the need for high-level annotation accompanying the documents to be identified. [17] and [18] provide the annotation with observed context hints of postings. Essentially, these are still supervised methods and there will still be problems when the text to be identified includes some languages which are not in training corpus.

For our requirement to purify a text where we have little linguistic knowledge of the language or languages present, this presents a problem and raises the key question: Can this annotation for identifying language be generated automatically and can unsupervised methods be used to identify language or at least classify a text into the different languages present.

In our own work we are attempting to fully automatically build the front-end language-processing component of a speech synthesis system. This component is required to take text in a given language, of which we have little of no linguistic knowledge, and produce a linguistic representation of the sounds and structure required to speak the text. We can achieve this using methods such as vector space models [19] but to do so we require pure data in a single language as input. As we usually dealing with low-resourced minority languages and the data we are using is often found data and we do not have the expertise to time to manually clean up data-sets. A typical scenario is that we wish to create a monolingual corpus from Wikipedia and similar web sites, the data crawled from these web sources is generally a mix of several languages either due to code-switching within the text of one language itself or due to the text having been partially translated from another language.

The existing supervised or bootstrapped approaches are unsuited to this problem and we require a completely unsupervised language identification method. [20] demonstrate an approach using similarity measures, but performance is greatly reduced when compared to supervised methods. [21] present a promising co-occurrence graph approach, but each language present must have a minimum of 100 sentences, which makes it unsuited to our particular task.

This paper addresses these issues and presents an unsupervised language identification method using the raw n-gram count to characterise features and a reformulated Latent Dirichlet Allocation (LDA) topic model. This approach is tested on the ECI/MCI Benchmark and a Wikipedia Swahili corpus and compared with other existing approaches. Additionally, we propose a new measure based on Minimal Description Length (MDL) to determine the number of languages present which we argue is more appropriate than the perplexity measure usually employed when modelling topics.

2. Latent Dirichlet Allocation for Language Identification

To be able to identify language in an unsupervised fashion we adopt and adapt a model from the field of Topic modelling. The most common topic model currently in use, is the generalisation of pLSI into Latent Dirichlet Allocation, which allows documents to contain a mixture of topics, developed by [22].

The basic idea behind traditional LDA is that documents are represented as random mixtures over latent topics, where each topic is characterised by a distribution over words [22]. To adapt this to language identification we consider sentences represented as random mixtures over latent languages, where each language is characterised as a distribution over letter n-grams counts. In such way, the document∼Language and Language∼N-gram hierarchies can similarly be modelled by the LDA for language identification, We call this approach LDA-LI for short. Algorithm 1 gives the pseudo code of generative LDA-LI model.

```
// Language plate
for all languages k∈[1, K] do
    sample components $\vec{\phi}_k \sim Dir(\beta)$;
end

// document plate
for all documents j∈ [1, D] do
    sample mixture proportion $\vec{\theta}_j \sim Dir(\alpha)$;
    sample document length $N_j \sim Poiss(\xi)$;

    // n-gram plate
    for all n-grams i∈ [1, Nj] in document do
        sample Language $z_{ij} \sim Mult(\vec{\theta}_j)$;
        sample N-gram $x_{ij} \sim Mult(\vec{\phi}_{z_{ij}})$;
    end
end
```

Algorithm 1: generative model of LDA-LI

2.1. Gibbs sampling and inference

To maintain the parallel with the topic modelling literature we will continue to discuss the model in terms of topics in document meaning languages present in sentences, only speaking specifically about language when it is directly appropriate.

The learning algorithm in this paper is based on the Collapsed Gibbs Sampling (CGS) [23], a Markov-chain Monte Carlo method. The model parameter $\phi = \{\vec{\phi}_k | \vec{\phi}_k \sim Dir(\beta)\}$, the set of topic distributions, can be integrated using the Dirichlet-multinomial conjugacy. The posterior distribution $P(Z|W)$ can then be estimated using the Collapsed Gibbs sampling algorithm, which, in each iteration, updates each topic assignment $z_{ij} \in Z$ by sampling the full conditional posterior distribution:

$$p(z_{ij} = k | Z_{\overline{ij}}, x_{ij} = w, W_{\overline{ij}}) \propto (C^{doc}_{kj} + \alpha) \frac{C^{word}_{kw} + \beta}{\sum\limits_{v'} C^{word}_{kv'} + W\beta} \qquad (1)$$

where $k \in [1, K]$ is a topic, $w \in [1, W]$ is a word in the vocabulary, x_{ij} denotes the i-th word in document j and z_{ij} the topic assigned to x_{ij}. $W_{\overline{ij}}$ denotes the words in the corpus with x_{ij} excluded, and $Z_{\overline{ij}}$ are the corresponding topic assignments of $W_{\overline{ij}}$. In addition, C^{word}_{kw} denotes the number of times that word w is assigned to topic k not including the current instance x_{ij} and z_{ij}, and C^{doc}_{kj} the number of times that topic k has occurred in document j not including x_{ij} and z_{ij}. Whenever z_{ij} is assigned to a sample drawn from (1), matrices C^{word} and C^{doc} are updated. After enough sampling iterations to burn in the Markov chain, $\theta = \{\vec{\theta}_j\}_{j=1}^{D}$ and $\phi = \{\vec{\phi}_k\}_{k=1}^{K}$ can be estimated by

$$\theta_{kj} = \frac{C^{doc}_{kj} + \alpha}{\sum\limits_{i=1}^{K} C^{doc}_{ij} + K\alpha}, \qquad (2)$$

$$\phi_{kv} = \frac{C^{word}_{kv} + \beta}{\sum\limits_{j=1}^{W} C^{word}_{kj} + W\beta} \qquad (3)$$

From equations 2 and 3, we see that the CGS learning and inference are some kinds of pseudo counts of original corpus. The implementation of CGS used in this paper is based upon the implementation of [24] using a Map-Reduce parallel framework with efficiency improvements by [25] using a Message Passing Interface(MPI)[1].

2.2. Feature space and model selection

One merit of LDA is that it inherently provides some degree of the automatic smoothing. [26] point out the LDA is a flexible latent variable framework for modeling sparse data in extremely high dimensional spaces. Even with the default hyper-parameter settings of those learning algorithms, LDA can smooth the sparse count data and infer on unseen data [22]. Thus in this paper we have just used the raw n-gram counts as the features.

The corpus is converted into samples by considering each individual sentence a document. These documents are then converted into character based n-gram counts (tokens for spaces, and beginning and end of sentence markers are included for each document). [27] show that for supervised learning $n \leq 3$ is sufficient n-gram length, but as we are attempting unsupervised learning, we include n-grams with n in the range 1-5 in an attempt to capture more information across both short and long contexts. Due to the smoothing ability of LDA to large sparse data discussed in section 2.1, we are able to use the raw n-gram counts without any smoothing, pruning or interpolation. In practice, the smoothing and pruning is actually realised by the hyper parameters α and β , which are configured with their default small values(<1) suggested by [25].

[1] https://code.google.com/p/plda/

An important issue with the LDA topic model is how to determine that an adequate number of individual topics are being modelled. In most cases [22, 23, 28, 29, 30], perplexity is used to evaluate the resulting model on held-out data. In our experiment, we found that the perplexity always reduces as the number of topics is increased, as shown in figure 9 of [22] and figure 10 of [28], this continues beyond the point where the number of topics in the model is equal to the actual number of different languages in the data.

To address this issue, we introduce a new measure base on the Minimal Description Length principle to find the smallest topic number with the best inference performance. MDL was introduced by [31]. Here we use the refined MDL formulations of [32],and normalise it by the joint entropy of a term (as opposed to the word) w in the test set and the given topic model T by

$$\begin{aligned} H(w,T) &= H(T) + H(w|T), \\ T &= \arg\min_{T} H(w,T) \end{aligned} \tag{4}$$

In this way, (4) punishes the model with more topics on the first term.

Evaluation of this MDL measure still ongoing but generally appears to work well. It is not reported further here as it is not of direct importance because for the task of text purification, where a majority language is present, we will show that if we set the number of languages to 2 then the system performs well and is able to separate out the non-majority languages into a single class.

3. Experimental evaluation

Table 1 summarises the performance of the LDA-LI model as a general unsupervised language identification tool, we find high precision and recall and when compared to other existing supervised systems our performance is not far behind. To evaluate the LDA-LI model specifically for the language purification task we perform the two experiments reported here.

In experiment 1 we try to simulate the typical scenario of unsupervised language filtering using the ECI/MCI[2] data, a benchmark corpora for language identification studies [33]. A majority language is mixed with an unknown number of other languages (with and without kinship) with difference ratios. We focus on the precision and recall of the majority language using our LDA-LI system.

In experiment 2 we investigate the ability of our model to filter a real Swahili corpora crawled from Wikipedia, where Swahili is the majority language in the content, but is mixed with additional material from an unknown number of other languages. Here we investigate the ability to filter this data and find pure Swahili sentences with the number of topics set to 2.

3.1. Experiment 1

In experiment 1, we simulate the typical scenario of unsupervised language filtering by mixing similar languages together in different proportions to determine how well the proposed system can identify and purify the primary language.

In the case1, German is the considered the principal language and mixed with Dutch, English and Turkish. We increasing add equal amounts of Dutch English and Turkish to lower the overall proportion of the primary language present.

We report the results for proportions of the other languages between 1 and 50% so that there is alway more German than the other languages combined so that German remains the majority language. In the case2, Dutch is the chosen as the major language but mixed with German, English and Turkish to provide a comparison.

For this experiment we fix the topic number of LDA-LI to 2, as in practice we would not know the number of other languages present and wish to apply unsupervised language filtering, without this knowledge. Figures 1 and 2 show the results for German and Dutch respectively. Where German is the primary language we see a very high precision whilst the ration of the other languages is less than 0.25 after which point the precision drops or becomes a little erratic. However the recall is lower and the model is quite conservative in identifying all of the German sentences.

Where Dutch is the primary language (Figure 2 we see a similar picture with precision, but here the recall is generally much higher, except when the mix of other languages present is less than 0.1, where we hypothesis there isn't sufficient data to have a good model of what the non-Dutch language class looks like.

In general we see that our LDA-LI system can purify the main language with high precision when it has $\leq 30\%$ other languages which is the typical scenario of language purifying.

We can get around the low recall since we can execute the LDA-LI one more time for each cluster of to improve the recall almost without degrading the precision. This kind of iterative use can also be used to improve the precision when mixed ratio is above 30%.

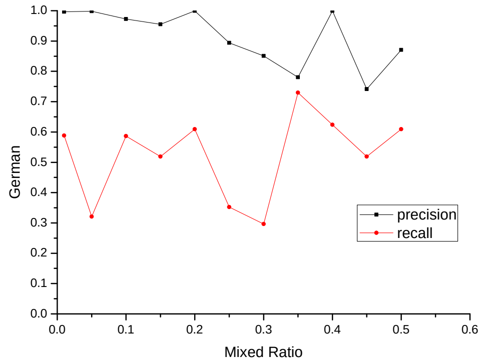

Figure 1: German purification result. German is mixed with different rations of other languages

3.2. Experiment 2

In experiment 2, we use a Swahili corpus crawled from Wikipedia. This nominally consists of 172,724 Swahili sentences but actually contains a mix of Swahili and other languages. Here, we also evaluate the LDA-LI model using 2 topics (i.e. Swahili and Other) as a filtering procedure to find pure Swahili sentences. Because it was unreasonable to determine the language or languages present in every sentences of this corpus manually, we used the *langID.py* with its pre-trained model to evaluate the performance of LDA-LI. That is, we take the inferences of *langID.py* as the underline correct language in the

[2] http://www.elsnet.org/eci.html

Method	sentence length of test	precision	recall	Fscore
LDA-LI(12)	Max 1297, Min 10 Average 81.65 characters	93.18%	92.97%	92.98%
langID.py		95.71%	96.00%	95.67%
Guesss_language		99.27%	95.00%	96.99%
ICF	100 characters	97.10%	97.50%	97.30%

Table 1: performance over 10-fold CV
Note: ICF method is from [10]

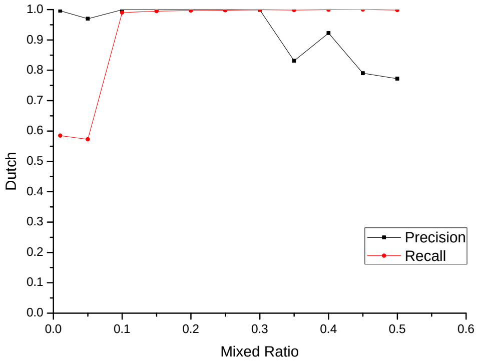

Figure 2: Dutch purification result. Dutch is mixed with different rations of other languages

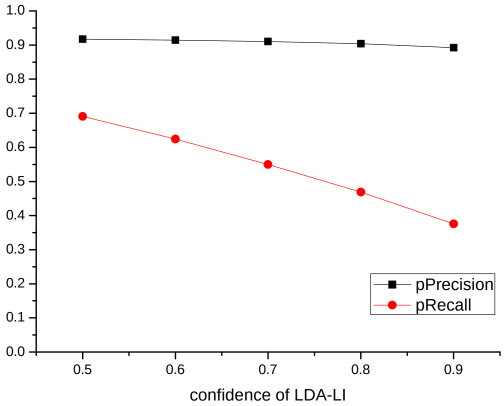

Figure 3: pPrecision and pRecall of LDA-LI when purifying a corpus of Swahili mixed other language data

corpus to compute pseudo precision and recall values (pPrecision and pRecall respectively in figure 3 for the LDA-LI model with the probabilities greater than thresholds between 50% and 90%. This probability measure is the confidence score from the LDA-LI, e.g., if the required probability is 90%, the LDA-LI only outputs all those sentences which are confirmed as Swahili with the probability$\geq$ 90%. Because we set the topic number fixed to 2, the smallest confidence is 50%.

We see that the changing the required probability of the LDA-LI from 50% to 90% has a negligible affect on the precision of LDA-LI while it degrades the recall. The reason for this is that in addition to any misinferences of LDA-LI, there are also many very short sentences, which are Swahili and correctly identified by LDA-LI but cannot be recognised by *LangID.py*. This means the actual precision will be higher than that shown in figure 3.

For the case where a sentences comprises multiple languages, we find in practice, that the presence of one or two other-language words is sufficient to classify the sentence as not being Swahili. This is the behaviour we desire for the task we are investigating, but may not always be appropriate.

4. Conclusion and future work

In this paper, we presented LDA-LI, an unsupervised language identification approach which takes raw 1-5 gram counts as features and allows us to both classify sentences by language or filter sentences not of the majority language from a corpora. We can identify the number of languages present by a measure base on the Minimal Description Length principle. And our experiments show that the LDA-LI is robust both for initial annotation of unknown languages and for further inferring and filtering.

For the language purification task we have shown that the LDA-LI system purifies with a high precision for mixes of languages similar to those we would require the task for. This makes it a useful tool for preparing found language corpora for building speech synthesis front ends, and for recording-script production in these languages, as a pure script is easier for a subject to record and provides a better training set for acoustic models.

As the primary language becomes less the majority language present precision begins to suffer. It is possible in this case that it would be better to use the system as a more general language identification tool and allow it to classify the individual non-majority languages present into their own categories as we have shown precision to remain high here.

The current system would reject sentences containing the primary language of interest mixed with another language (in the same sentence) but the LDA framework allows for multiple topics to be assigned, so there is scope for further investigation here.

5. Acknowledgments

This work is supported by Simple4All: funded by the European Community's Seventh Framework Programme (FP7/2007-2013)under grant agreement no.287678 and the Chinese Scholarship Council.

6. References

[1] T. Dunning, *Statistical identification of language*. Las Cruces. New Mexico: Computing Research Laboratory. New Mexico State University, 1994.

[2] A. Xafopoulos, C. Kotropoulos, G. Almpanidis, and I. Pitas, "Language identification in web documents using discrete hmms," *Pattern Recognition*, vol. 37, no. 3, pp. 583–594, 2004.

[3] P. Sibun and J. C. Reynar, "Language identification: Examining the issues," 1996.

[4] W. J. Teahan, "Text classification and segmentation using minimum cross-entropy," 2000.

[5] J. Hakkinen and J. Tian, "N-gram and decision tree based language identification for written words," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, pp. 335–338.

[6] J. Tian and J. Suontausta, "Scalable neural network based language identification from written text," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–48–51 vol. 1.

[7] L.-F. Zhai, M.-h. Siu, X. Yang, and H. Gish, "Discriminatively trained language models using support vector machines for language identification," in *The Speaker and Language Recognition Workshop, Odyssey 2006*. IEEE, 2006, pp. 1–6.

[8] K. N. Murthy and G. B. Kumar, "Language identification from small text samples," *Journal of Quantitative Linguistics*, vol. 13, no. 1, pp. 57–80, 2006.

[9] C. Kruengkrai, P. Srichaivattana, V. Sornlertlamvanich, and H. Isahara, "Language identification based on string kernels," in *Communications and Information Technology, 2005. ISCIT 2005. IEEE International Symposium on*, vol. 2. IEEE, 2005, pp. 926–929.

[10] H. Takçı and T. Güngör, "A high performance centroid-based classification approach for language identification," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2077–2084, 2012.

[11] B. King and S. Abney, "Labeling the languages of words in mixed-language documents using weakly supervised methods," in *Proceedings of NAACL-HLT*, 2013, pp. 1110–1119.

[12] H. Yamaguchi and K. Tanaka-Ishii, "Text segmentation by language using minimum description length," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 969–978.

[13] T. Baldwin and M. Lui, "Language identification: The long and the short of the matter," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 229–237.

[14] H. Romsdorfer and B. Pfister, "Text analysis and language identification for polyglot text-to-speech synthesis," *Speech Communication*, vol. 49, no. 9, pp. 697–724, 2007.

[15] M. Lui and T. Baldwin, "langid. py: An off-the-shelf language identification tool," in *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, 2012, pp. 25–30.

[16] A. Kilgarriff and G. Grefenstette, "Introduction to the special issue on the web as corpus," *Computational linguistics*, vol. 29, no. 3, pp. 333–347, 2003.

[17] U. F. Mayer, "Bootstrapped language identification for multi-site internet domains," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 579–585.

[18] M. Goldszmidt, M. Najork, and S. Paparizos, *Boot-Strapping Language Identifiers for Short Colloquial Postings*. Springer, 2013, pp. 95–111.

[19] O. Watts, A. Stan, R. Clark, Y. Mamiya, M. Giurgiu, J. Yamagishi, and S. King, "Unsupervised and lightly supervised learning for rapid construction of tts systems in multiple languages from 'found' data: evaluation and analysis," in *Proc. 8th ISCA Speech Synthesis Workshop*, 2013. [Online]. Available: http://consortium.simple4all.org/files/2013/07/watts_SSW8_2013.pdf

[20] A. Amine, Z. Elberrichi, and M. Simonet, "Automatic language identification: An alternative unsupervised approach using a new hybrid algorithm." *IJCSA*, vol. 7, no. 1, pp. 94–107, 2010.

[21] C. Biemann and S. Teresniak, "Disentangling from babylonian confusion–unsupervised language identification," in *Computational Linguistics and Intelligent Text Processing*. Springer, 2005, pp. 773–784.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[23] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[24] Y. Wang, H. Bai, M. Stanton, W.-Y. Chen, and E. Y. Chang, *Plda: Parallel latent dirichlet allocation for large-scale applications*. Springer, 2009, pp. 301–314.

[25] Z. Liu, Y. Zhang, E. Y. Chang, and M. Sun, "Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 26, 2011.

[26] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On smoothing and inference for topic models," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 27–34.

[27] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," *Ann Arbor MI*, vol. 48113, no. 2, pp. 161–175, 1994.

[28] D. Newman, A. Asuncion, P. Smyth, and M. Welling, "Distributed algorithms for topic models," *The Journal of Machine Learning Research*, vol. 10, pp. 1801–1828, 2009.

[29] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1105–1112.

[30] B. Grün and K. Hornik, "topicmodels: An r package for fitting topic models," *Journal of Statistical Software*, vol. 40, no. 13, pp. 1–30, 2011.

[31] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[32] P. Grünwald, "A tutorial introduction to the minimum description length principle," 2005.

[33] S. Armstrong-Warwick, S. Thompson, D. McKelvie, and D. Petitpierre, "Data in your language: the eci multilingual corpus 1," in *Proc. International Workshop on Sharable Natural Language Resources, Japan*. Citeseer, 1994, pp. 97–106.