Inferring Phylogenetic Graphs of Natural Languages using Minimum Message Length

Jane N. Ooi and David L. Dowe

School of Computer Science and Software Engineering, Monash University,
Clayton, Vic 3800, Australia
janeo@bruce.csse.monash.edu.au

Abstract. We extend phylogenetic (or evolutionary) trees to phylogenetic graphs. Unlike phylogenetic trees, phylogenetic graphs are capable of modelling evolution where a child node inherits from more than one parent node. Minimum Message Length (MML)(Wallace and Boulton 1968; Wallace 2005) is an inductive inference method that measures the goodness of a model. We use MML to infer phylogenetic graphs (including mutation probabilities along arcs). We introduce the use of MML to infer phylogenetic graphs for artificial languages as well as for some European languages (English, French and Spanish). Our modelling assumes only copy and change operations on characters, and is based on words which have the same length in all natural languages considered.

1 Introduction

Evolution of languages happens gradually around us everyday. As modernisation of society takes place, new words and new grammatical structures are created or adapted from some languages into different languages. Our aim is to be able to model this evolution and describe the relationships between different languages.

A phylogenetic model shows the evolutionary interrelationship among various species or other entities. In this article, we initially consider a phylogenetic model of natural languages as an evolutionary tree that shows how different languages have descended and evolved from one another. We then generalise this by introducing the notion of phylogenetic graphs, which are like phylogenetic trees but they permit nodes to have more than one parent. Whereas nodes in a phylogenetic tree (other than the root node) must have one common ancestor, this is not necessarily true of *phylogenetic graphs*. We then apply these techniques to natural language text. The languages that will be used include artificial languages and some European languages (English, French and Castillian Spanish). Words have been chosen which have the same lengths in all languages, as our preliminary model assumes only copy and change operations on characters. Accents on characters have been ignored. (This paper is expanded in [10].)

2 Language Compression in building phylogenetic trees

Many previous works inferring phylogenetic trees for languages have been carried out using language compression techniques.

In [4], thirty-three versions of a chain letter (from between 1980 and 1995) were collected. The measure of similarity between these chain letters is estimated by compressing the chain letters two at a time. Chain letters that are similar to each other produce a smaller compression size. From the results of comparing chain letters, a phylogenetic tree was inferred. The resulting tree appears to be a "perfect" phylogeny [4], where letters that share the same characteristic are always grouped together. In earlier work [3], a similar method of comparing languages used the Lempel and Ziv algorithm (LZ77) [19] to compress languages. The relative entropy between languages was calculated, as languages with lower relative entropy have more similarities between them. Using this method, the authors created a language tree by comparing the translations of "The Universal Declaration of Human Rights" in over 50 languages [3].

Generalising and allowing a language to have more than one parent yields a phylogenetic graph rather than a tree structure. We will use Minimum Mesage Length (see section 3) to infer these, starting in section 4.

3 Minimum Message Length (MML)

We use the information-theoretic Minimum Message Length (MML) [15, 18, 16, 14] principle here to infer phylogenetic trees for languages largely because of its theoretical optimality properties and its wide-ranging achievements in a vast range of inference problems - see, e.g., [16, 7, 6, 17, 13, 14].

MML encodes a body of data as a two-part message. The first part consists of the hypothesis about the data. The second part is the optimal encoding of the data given that the hypothesis stated in the first part is true. Hence, the message length for data encoded using MML would be

$$MsgLength = MsgLength(Hypotheses) + MsgLength(Data|Hypotheses)$$

If we have a good hypothesis about the data, we save a lot of space in encoding the data. MML states that the best encoding of the data would be the one which produces the smallest two-part message length. For discussions of the relationship between MML, the works of Solomonoff [12], Kolmogorov [9] and Chaitin[5] (and the subsequent Minimum Description Length (MDL) principle [11]) see, e.g., Wallace and Dowe [16], Comley and Dowe [7] and Wallace [14].

Allison, Wallace and Yee [2] have previously applied MML methods to infer evolutionary trees for DNA sequences. They used MML to calculate the posterior odds-ratio of two competing phylogenetic trees' hypotheses. A finite-state machine is used to model the mutation process between DNA sequences. In this article, we use MML algorithms to compress the vocabularies of languages for comparing the similarities between them.

3.1 Multi-state message length and Parameter estimation

The MML parameter estimation for a discrete multi-state distribution discussed in [17] will be used to model the mutation between languages.

For a multi-state distribution with M states, a uniform prior, $h(\mathbf{p}) = (M-1)!$ is assumed over the $(M-1)$-dimensional region of hyper-volume $1/(M-1)!$ given by $p_1 + p_2 + ... + p_M = 1; p_i \geq 0$. The parameters for each state are estimated as given by [15, p187(4), p194(28), p186(2)][13, sec. 5.1][17, eq. 5]

$$\hat{p}_m = \frac{n_m + 1/2}{N + M/2}$$

where n_m is the number of things in state m and $N = n_1 + n_2 + ... + n_M$. These parameter estimates lead to the message length being minimized.

Calculating the overall message length for stating both the parameters and the data encoded using these estimated parameters is (correcting a typo in [17, eq. 6])

$$\frac{M-1}{2}(\log(\frac{N}{12}) + 1) - \log(M-1)! - \Sigma_{m=1}^{M}(n_m + 1/2)\log\hat{p}_m$$

4 Building a phylogenetic model

To build a phylogenetic model of various languages, the vocabularies of these languages must firstly be extracted. These vocabularies can then be compressed using Minimum Message Length (MML) methods (recall sec. 3). The similarity of language A with languages B,C,D... can be compared by firstly compressing language A alone, noting the size of the compression. Next, languages B,C,D... are appended to language A one at a time and the compressor compresses these using a model of their relation to language A. The compressed file size is observed and compared to the file size that was previously obtained without reference to language A. Languages that have many similarities with language A would produce a smaller compressed file size as compared to languages that are totally different from language A.

Using the method mentioned above, we are then able to compare the similarities between languages.

4.1 Tree and Graph topologies

We will be using 3 languages and considering 5 different topologies for them. They are as below:

Tree topologies

- Topology 1: The null hypothesis which assumes that all languages are unrelated.

```
language1   language2   language3
```

- Topology 2: The topology assuming that only 2 out of the 3 languages are related.

```
language1   language2 -> language3
```

- Topology 3: The tree topology assuming that children language 2 and language 3 descend from language 1.

```
        language1
         /      \
        v        v
language2     language3
```

Graph topologies

- Topology 4: The graph topology assuming that language 3 descends from parents language 1 and language 2.

```
language1     language2
         \       /
          v     v
        language3
```

- Topology 5: The topology assuming that language 2 descends from language 1, and that language 3 descends from parents language 1 and language 2. (Note, though, that the copy/change mutation relation between languages 1 and 2 is symmetric.)

```
language1  -> language2
         \       /
          v     v
        language3
```

4.2 MML method of costing tree and graph topologies

The method of costing MML decision graphs [13] and Generalised Directed Acyclic Bayesian Networks (which deal with a hybrid mix of continuous and discrete variables) [6, 7] will be adapted to cost the phylogenetic graphs.

We assume uniform prior probabilities for each of the 5 topologies from section 4.1. Hence it will cost $-\log(1/5)$ to encode a particular topology. The root language is encoded with all characters costing $\log(26+1) = \log(27)$, ignoring frequencies and not using a multinomial message length. Recalling multinomial message lengths from sec. 3.1, a child language is encoded using binomial (for a tree) or multinomial (for a graph) mutations. This is so because our simple model assumes only copy and change (and neither insert nor delete) operations on characters. Detailed costing of each topology in section 4.1 is discussed below:

Encoding Tree topologies

- Topology 1: Each language is costed separately, costing $\log(27)$ per character.
- Topology 2: Language 1 is costed separately. An extra cost of $-\log(1/3)$ is needed to determine which language sits in the root node of the tree. The parent language is encoded, and the child language is encoded in terms of the parent language.
- Topology 3: An extra cost of $-\log(1/3)$ is needed to determine which language sits in the root node of the tree. The parent language is encoded, and then each child language is encoded in terms of the root parent language.

Graph topologies

- Topology 4: An extra cost of $-\log(1/3)$ is needed to determine which language sits in the child node of the tree.
  We have considered and used 2 possible ways of encoding the child language in terms of the parent languages :
  - [Topology 4a]: Both parent languages are encoded, and then the child language is encoded as a discrete trinomial distribution stating for each character where it descends from (parent 1, parent 2 and not parent 1, or is a new character).
  - [Topology 4b]: Both parent languages are encoded, and then the child language is encoded as a discrete multi-state distribution for each character depending on whether parent characters agree or disagree. If both parents agree, state whether child character agrees or disagrees. If parents disagree, state (using a trinomial distribution) whether child character comes from parent 1, parent 2 or is a new character.
- Topology 5 [Topologies 5a and 5b]: This topology is encoded the same as topology 4 except that parent 2 is encoded in terms of parent 1. (Because of symmetry, an additional cost of $-\log(1/2)$ should not actually be required to determine which language is parent 2.)

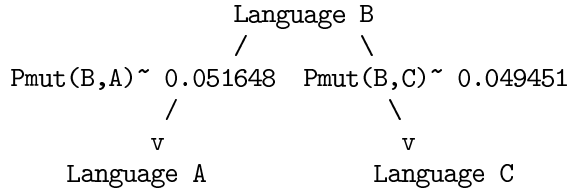
4.3 Encoding descended languages

Each child node is encoded as a discrete multi-state distribution. Using MML parameter estimation, we infer a probability of mutation p_m from each parent language to each child language. While encoding the data, we then state for each character of the child language whether it is similar to the corresponding character of the parent language, or it is a mutation. In the case of a mutation, we then have to send the new character following it. This costs $\log(Z-1) = \log(26-1) = \log(25)$, as we now know that the character is not the same as that of its parent language.

5 Phylogenetic tree for artificial languages

To test the method we have discussed, 3 sets of vocabularies (50 words each) of artificial languages are created. A subset of each vocabulary is shown in Table 1. Set A is totally random, consisting of 27 characters (A-Z and .). Set B is 5% mutated from Set A, and Set C is 5% mutated from Set B. Each word has the same length as its corresponding ("translation") mutation in all 3 sets. Using the tree inference program we have produced using the abovementioned methods, we infer the phylogenetic tree for these 3 vocabularies.

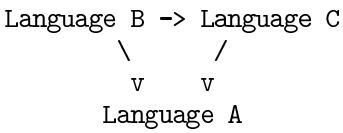
Our results correctly show a tree using Topology 3 where language B is the root language and language A and language C descend from it.



This is the expected result as we have created vocabularies of equal length. Hence a mutation from language A to language B is equivalent to a mutation from language B to language A. The MML inferred probability of mutation between language A and language B is 0.051648, whereas the MML inferred probability of mutation between language B and language C is 0.049451. This refers to character-to-character mutation and is very close to the actual mutation probabilities (both 0.05) used to generate these languages.

Detailed cost of tree:
Cost of parent language (Lang. B) (no. of chars * log(27)) = 2158.718926 bits
Cost of child language (Lang. A) binomial distribution = 392.069784 bits
Cost of child language (Lang. C) binomial distribution = 378.562159 bits
Total tree cost = log(5) + log(3) + 2158.72 + 392.07 + 378.56
= 2933.257759 bits

The cheapest phylogenetic graph is a graph (using topology 5b) where Language B is the parent node of language C, and language A is the child node of both language B and language C.



MML inferred probability of mutation between language B and C = 0.049451
Cost of binomial distribution, (language B − > language C) = 378.562159 bits

MML inferred probability that both parents agree = 0.950549
-MML inferred probability that child(A) agrees with both parents = 0.903297
-MML inferred probability that child(B) disagrees with parents = 0.096703
Cost of this binomial distribution = 381.299676 bits
MML inferred probability that both parents (B and C) disagree = 0.049451
-MML inferred probability of coming from parent y = 0.891304
-MML inferred probability of coming from parent z = 0.065217
-MML inferred probability that does not come from parents = 0.043478
Cost of this trinomial distribution = 38.579308 bits

Total message length = 2962.066959 bits

As language B and language C are very similar with only 5% mutation and both languages can be thought of as either directly or indirectly descending from language A, such a graph is not unexpected in our findings.

6 Phylogenetic graph for European languages

With the satisfactory results obtained in sec. 5 from artificial languages, we now move on to European languages. We chose English, French and (Castillian) Spanish. We selected a vocabulary of 30 words for each of these 3 languages from [8] (which was only available in printed hard copy). Accents on characters have been removed and, because our preliminary model uses only copy and change operations (and no insert and no delete operations) on characters, each word has the same length as its corresponding translation in all 3 sets. Table 2 shows the list of words we have used.

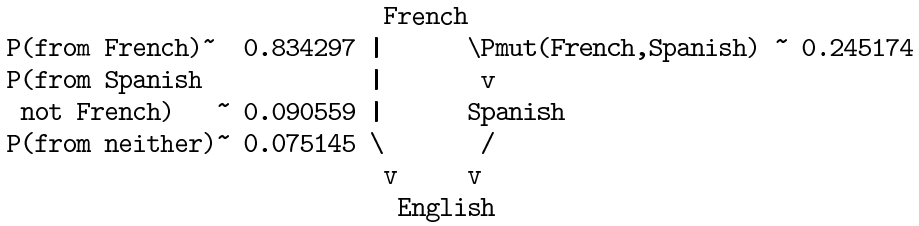
Lang. A	Lang. B	Lang. C
aera.	aera.	aera.
aertadaer.	aerradaer.	awrradaer.
aerya.	aerya.	aerva.
air.	air.	afr.
asdfge.	assfge.	assfge.
asrpyas.	asrpyas.	asrpyas.
asrtma.	asrtma.	tsrtma.
astakera.	astakera.	astakera.
awefadfger.	awefabfger.	awefabfger.
awet.	awet.	awet.
bser.	bser.	bher.
bsoty.	bsoty.	bsoty.
bsrtyaj.	bsttyaj.	bstteaj.
dfddgr.	dfddtr.	dfddtr.
dfg.	vfg.	vfg.
dfpelmy.	dfpelmy.	dfprlmy.
eer.	eer.	eer.
ert.	ert.	ert.
ewg.	ewg.	ewg.
gaerd.	gserd.	gserd.
gijs.	gijs.	gijh.
hdcvsery.	hdcvsery.	hdcvsery.
hgsryujk.	hgsryujk.	hgsryujk.
hyergf.	hytrgf.	hetrrf.
kioln.	kioln.	kiohn.
mqzo.	mqeo.	mqeo.
pdfb.	pdfb.	pdfb.
qpwomz.	qpwtmz.	qpatmz.
qvery.	qvery.	qvery.
zlsdrya.	zlcdrya.	zlchrya.

Table 1: Artificial Languages (sec. 5)

English	French	Spanish
baby	bebe	nene
beach	plage	playa
biscuits	biscuits	bizcocho
camping	camping	camping
cabaret	cabaret	cabaret
centimetres	centimetres	centimetros
cream	creme	crema
disaster	desastre	desastre
europe	europe	europa
excursion	excursion	excursion
facial	facial	facial
jack	cric	gato
jumper	jumper	jersey
kilometres	kilometres	kilometros
litres	litres	litros
lottery	loterie	loteria
opera	opera	opera
overseas	outremer	exterior
patisserie	patisserie	confiteria
reception	reception	recepcion
sauna	sauna	sauna
service	service	servido
spade	pique	pique
stop	stop	pare
souvenir	souvenir	recuerdo
taxi	taxi	taxi
vinegar	vinagre	vinagre
waitress	serveuse	camarera
young	jeune	joven
zero	zero	cero

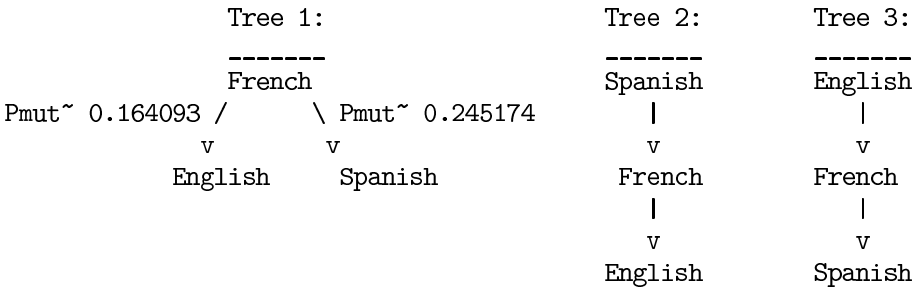
Table 2: European Languages (sec. 6)

Using our inference method above, the best model we achieved with this limited size vocabulary is the phylogenetic graph using topology 5a where Spanish is related to French (recalling secs. 4.1 and 4.2), and English descends from both French and Spanish.



Detailed cost of graph:
Cost of parent language (French) (no. of chars * log(27)) = 1226.760976 bits
Cost of parent/"child" language (Spanish) binomial distribution = 734.59 bits
Cost of child language (English) trinomial distribution = 537.698815 bits
Total tree cost = log(5) + log(3) + log(2) + 1226.76 + 734.59 + 537.70
= 2503.954019 bits

The closest tree topology shows French as the root parent with English and Spanish both descending from French. However, as our vocabularies contain words of similar length, we can conclude that a mutation from string A to string B is equivalent to a mutation from string B to string A. Hence with the results obtained, we know that there are 3 equivalent possible trees that can be concluded from Topology 3. They are:



Detailed cost of tree:
Cost of parent language (French) (no. of chars * log(27)) = 1226.76 bits
Cost of French $->$ child (Spanish) binomial distribution = 734.59 bits
Cost of French $->$ child (English) binomial distribution = 549.88 bits
Total tree cost = log(5) + log(3) + 1226.77 + 734.59 + 549.88 = 2515.1 bits

7 Conclusion and future work

We have used a simple MML model to infer phylogenetic trees and (our new and more general notion of) phylogenetic graphs for both artificial languages and some European languages (English, French and Spanish) by ignoring accents and assuming a limited model of copy/change evolution which requires all compared words to have the same length. We were able to verify our methods of inferring mutation probabilities between languages using artificial languages with known mutation probabilities. In work in progress (see, e.g., [10]), we are using string alignment techniques [1] and finite state machines as studied in [2] to model inserts, deletes and words evolving to a different length. We also plan to infer phylogenetic trees of languages using their grammatical structure as well as their vocabularies. We further aim to refine our methods and use this to study the endangered languages of the Aboriginal peoples of Australia.

References

1. L. Allison and C.S. Wallace. An information measure for the string to string correction problem with applications. *17th Australian Comp. Sci. Conf.*, pages 659–668, Jan 1994.
2. L. Allison, C.S. Wallace, and C.N. Yee. Minimum message length encoding, evolutionary trees and multiple-alignment. *Hawaii International Conference on System Science*, 25:663–674, 1992.
3. D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. *Physical Review Letters*, 88, 2002.
4. C.H. Bennett, Ming Li, and B. Ma. Chain letters and evolutionary histories. *Scientific American*, 64-69:987–1006, 2003.
5. G.J. Chaitin. On the length of programs for computing finite sequences. *Journal of the Association for Computing Machinery*, 13:547–549, 1966.
6. Joshua W. Comley and D.L. Dowe. General Bayesian networks and asymmetric languages. *Proc. 2nd Hawaii International Conference on Statistics and Related Fields*, June 2003.
7. Joshua W. Comley and D.L. Dowe. Minimum Message Length, MDL and generalised Bayesian networks with asymmetric languages. *Advances in Minimum Description Length: Theory and Applications*, 11:267–294, April 2005. Final camera ready copy was submitted in October 2003.
8. Lixi Darvall. *The Illustrated International Phrase Book*. Julian Friedmann Publishers Ltd. London, 1979. ISBN 0-8317-4905-9.
9. A.N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7, 1965.
10. J.N. Ooi and D.L. Dowe. Inferring phylogenetic graphs for natural languages using MML. Technical Report 2005/178, Oct 2005. School of Computer Science and Software Engineering, Monash University, Clayton, Australia.
11. J.J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.
12. R.J. Solomonoff. A formal theory of inductive inference. *Information and Control*, 7:1–22,224–254, 1964.
13. P.J. Tan and D.L. Dowe. MML inference of Decision Graphs with Multi-Way Joins and Dynamic Attributes. *Proc. 16th Australian Joint Conf. Artificial Intelligence (AI'03), Perth, Australia, Springer LNAI 2903*, pages 269–281, Dec 2003.
14. C.S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Springer, 2005. ISBN 0-387-23795-X.
15. C.S. Wallace and D.M. Boulton. An information measure for classification. *Computer Journal*, 11:185–194, 1968.
16. C.S. Wallace and D.L. Dowe. Minimum message length and Kolmogorov complexity. *The Computer Journal*, 42(4):270–283, 1999.
17. C.S. Wallace and D.L. Dowe. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing*, 10:73–83, Jan. 2000.
18. C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society*, 49:240–265, 1987.
19. Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, IT-23(3), 1977.