Estimating the Cost of Adapting the Resources of One Language for Another

Anil Kumar Singh, Kiran Pala and Harshit Surana
Language Tech. Research Centre
IIIT, Hyderabad, India
anil@research.iiit.ac.in, kiranap@gmail.com, surana.h@gmail.com

Abstract

Developing resources which can be used for Natural Language Processing is an extremely difficult task for any language, but is even more so for less privileged (or less computerized) languages. One way to overcome this difficulty is to adapt the resources of a linguistically close resource rich language. In this paper we discuss how the cost of such adaptation can be estimated using subjective and objective measures of linguistic similarity for allocating financial resources, time, manpower etc.

1 Introduction

In a linguistically dense and diverse area like South Asia, the number of languages with a large number of speakers is quite high. At the same time, the resources for these languages are very scarce, and so are other resources (finance, time, manpower etc.) for creating these language resources. However, there is one fact which can make the task of building resources for these languages somewhat easier. This fact is the similarity of languages belonging to certain groups or families, partly the result of long historical proximity (Emeneau, 1956).

In such a situation, the ability to quantitatively measure the similarities and differences among languages as well as dialects (Dyen et al., 1992; Ellison and Kirby, 2006) can be very important, not only for providing evidence for or against a variety being a language or a dialect, but also for the more practical purpose of building resources for all these varieties, especially those which are relatively less privileged.

When two languages (or varieties) are quite close and one of them happens to be a more privileged one in the sense of having language resources, it may be possible to adapt the resources of the more privileged variety for the less privileged one with much less effort than would be required if those resources were to be created from scratch. Of course, not all resources can be adaptable. In the South Asian context, it is quite likely that a more privileged variety will be close to more than one less privileged variety. In such a case, the more privileged variety can be treated as a 'source' language or variety around which the development of the resources for a number of varieties can be centered. As an example, Hindi can be the source variety for Braj, Rajasthani, Avadhi, Bhojpuri (which are *considered* to be the dialects of Hindi) and perhaps even for Punjabi (which is *considered* to be a separate language).

In this paper we present the results of some experiments in trying to estimate the cost of developing resources for less privileged languages provided that there is a source language with some existing resources. We take some such source and some less privileged languages and try to estimate the cost in terms of different kinds of linguistic distances. We also discuss how such estimation of cost can be performed in a more subjective way, using the knowledge about the linguistic characteristics of the languages being considered.

2 Similarities of Languages

The way we study the differences and the similarities among languages depends on our purpose. So, for example, if we want to construct a genealogy of world's languages (Nerbonne, 2005), we might be able to achieve our goals even if restrict our study to phonology and lexicon. On the other hand, the study of typology and universals (McMahon and McMahon, 2005) will require us to cover other linguistic levels such as syntax and semantics. Similarly, when we want to estimate the cost of adapting resources of one language for another, our method for estimation will depend on the kind of resource we are trying to adapt. Strictly speaking, this is true. However, there is one interesting question: does the difference among language at one level (say, phonology) roughly also give an estimate of difference at another level (say, syntax)? The argument could be that if two languages are genetically distant (as found by studies at the levels of phonology and lexicon), they are also likely to be distant at the syntactic level because, in such a case, syntactic similarities can only be accidental. This argument implies that it is reasonable to assume that the distance among languages calculated on the basis of study at one or two levels roughly generalizes to other levels too.

3 Adaptable Resources

It is obvious that not all resources can be adapted even for very close languages. For example, Hindi and Urdu are close enough to be considered the same language, but resources like spell checker and tokenizer cannot be so easily adapted because the two languages use very different scripts.

Which resources can be adapted depends upon the kinds of similarities among the languages. Obviously, if two languages have similar morphology, a morphological analyzer developed for one can be adapted for the other. And if both the languages are free word order and have the same default word order (say, SOV) and also use post-positions to mark relations, then the parser built for one can possibly be adapted for the other. This is the case with languages like Hindi and Kannada, even though they belong to different linguistic families.

4 Cost of Resource Adaptation

Based on the discussion in the preceding sections, a strong claim can be made that measures of distances at a particular linguistic level can give us a reasonably good estimate of adapting a resource only if the linguistic level is relevant for that resource. A weak claim can also be made that the distance at a particular linguistic level can give us a rough estimate of the cost of adapting resources, in most cases irrespective of the relevant linguistic level.

The cost of adaptation also depends on the kind of resource, not just with respect to the relevant linguistic levels, but also on other characteristics of the resource. We will not be studying this second aspect of the cost of adaption. Therefore, once some estimate has been made of the cost of adaption based on the calculation of linguistic distance, the estimate may have to be revised to take care of other specific factors for the resource being adapted. Some of these factors might be completely non-linguistic, e.g. the practical constraints under which the resource developers will be working.

5 Objective and Subjective Estimation

As mentioned earlier, estimation of language distances can be either objective or subjective. The latter is needed when the former is not feasible. By objective we mean calculating the distance based on actual data (corpus) and using some computational technique without human involvement except in designing the method for computation. Subjective estimation, on the other hand, implies some human involvement in assigning numerical values, even during the process of estimation, not just while designing the method. The advantages and disadvantages of ‘objective’ and ‘subjective’ methods for our purpose are similar to those for other problem in computational linguistics, i.e., of using almost purely statistical method versus using almost purely linguistic methods.

We will present examples of both objective and subjective estimation in the following sections.

5.1 Objective Estimation

In this section we describe a method for objective estimation and present the results of some experiments using this method for many Indian languages or varieties. We will rely mainly on a corpus and the method itself would use linguistically informed measures of similarity.

5.1.1 Similarity Measures

We use two similarity measures for estimating the distance between languages. Both of these are designed to use some linguistic information at the level of writing system, phonology and lexicon (Singh and Surana, 2007). The basic idea is to first extract a list of highly frequent words from the unannotated corpus of each language. No other language resource is used. Since these words are highly frequent, they are likely to be from the core vocabulary of the language. This is in line with the insights gained from the techniques used in historical linguistics for comparing languages to find out whether they are related or not.

Then we use a method based on Computational Phonetic Model of Scripts (CPMS) to identify the cognates among these languages by calculating the surface similarity scores of pairs of words from the word lists extracted (Singh, 2006). By aligning word pairs using these scores and then using a threshold, we can identify the likely cognates, or more accurately, words of common origin since the method does not distinguish commonly inherited words from borrowed words.

The first measure of language distance (or of the cost of adaption of resources) is simply based on the idea that the more words of common origin the two languages have, the more likely they are to be similar. This measure is called the Cognate Coverage Distance (CCD). Cognate coverage distance gives us a measure of similarity of two languages, but it does so without taking into account the phonetic difference between two cognates. To include this factor, we use another measure called the Phonetic Distance of Cognates (PDC).

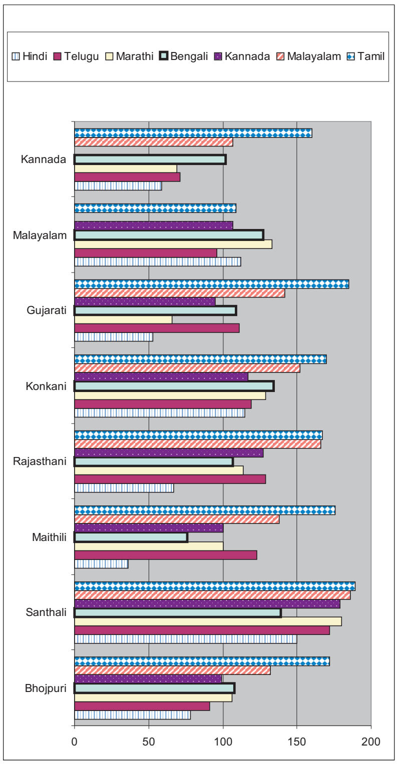

Figure 1: Selecting the best source language for some resource scarce languages. The length of the bar represents the distance and the cost of adaptation. The pattern in the bars represents a language.

5.2 Experiments

We conducted experiments on 17 varieties (languages or dialects). Out of these, four (Hindi, Bengali, Telugu and Marathi) are considered only as possible source (resource rich) languages. Three (Tamil, Malayalam and Kannada) are considered both as possible source languages as well as resource scarce languages. The reason for this is that these languages do not easily fit in either category: they have some resources, but they lack many others.

Ten varieties (Assamese, Bhojpuri, Gujarati, Konkani, Bishnupriya Manipuri, Maithili, Oriya, Rajasthani, Punjabi and Santali) are considered only as resource scarce languages as they hardly have any language resources. We had difficulty even in collecting unannotated corpus for many of the resource scarce languages like Maithili, Bhojpuri, Santali etc. Then there is the problem of encodings or notations in which the text is available. We also had to build encoding converters by manually preparing mappings for some of the languages. The text in Konkani was in an ASCII based notation, while that in Santali was in a different script called Ol Chiki.

For all the 17 varieties, we converted the text into UTF-8 format so that the CPMS based method could be applied and distances could be calculated using the measures (CCD and PDC) described earlier.

5.3 Results

The results are shown in Figure-1, which gives the distances between the source languages and the resource poor languages based on the measures CCD and PDC.[1]

As Figure-1 shows, most of the results are intuitively correct. For example, Hindi is found to be the source best language for Bhojpuri, Maithili and Rajasthani. All these three varieties are considered to be dialects of Hindi. Note that the purpose of these experiments is not just to find out the best source languages, but also to get a quantitative estimate of the cost of adaption of resources for the purpose of allocating finance, manpower, time etc. Another example is that Santali is shown to be closest to Bengali, although quantitatively it is not very close even to Bengali (as it belongs to a completely different language family than the other varieties).

6 Subjective Estimation

In subjective estimation, we may not use any measures that are applied on corpus. Instead, we do the following:

1. Select some linguistic features which are relevant to the resources being adapted.

2. For each language (or variety), assign some numerical value to the feature.

3. Calculate the distance between two languages using the numerical values of the feature. Different features can be given different weights depending upon their relevance.

The simplest way will be to select only boolean features or transform linguistic properties into boolean features. In such a case, assigning a numerical value (0 or 1) will be straightforward and so will be the calculation of distance. It is not required that all (or most of) the properties of languages be considered. We just have to pick up some representative and relevant (for resources) properties. There should be enough of them to give us a good estimate of the cost of adapting resources.

6.1 An Example

To demonstrate how subjective estimation can be performed, we will consider a hypothetical case where the resources we are interested in adapting are rule based morphological analyzer and parser. We assume that these are available for Telugu. The purpose is to build a parser for the other three South Indian (Dravidian) languages,

[1] The extended version of the paper, which contains more details about the results, is available at http://ltrc.iiit.ac.in/anil/papers/resource-adaptation.pdf

namely Kannada, Tamil and Malayalam. We wish to calculate the relative costs of adapting the Telugu resources for these three languages.

In the first step, we define the following features, mostly based on Caldwell (Caldwell, 1913), principally because they represent relevant characteristics and can also be given numerical values:

1. **DAG**: Degree of agglutination
2. **DIM**: Degree of inflection with respect to number, person and gender
3. **DIT**: Degree of inflection with respect to tense, aspect and modality
4. **FWO**: The extent to which free word order applies
5. **DWS**: Degree of word segmentation
6. A set of features in which each (boolean) feature represents the presence of a particular case: nominative (**CNT**), accusative (**CAT**), dative (**CDT**), sociative (**CST**), genitive (**CGT**), instrumental (**CIM**), locative (**CLT**), vocative (**CVT**) and ablative (**CAB**)
7. A set of (boolean) features for the types of pronouns: personal (**PPN**), adjectival (**PAV**) and relative (**PRT**)

	Weight	KN	ML	TM	TL
DAG	*3*	2	2	2	3
DIM	*3*	2	0	2	2
FWO	*4*	3	3	2	3
DWS	*3*	1	4	3	1
CNT	*2*	1	1	1	1
CAT	*2*	1	1	1	1
CDT	*2*	1	1	1	0
CST	*2*	0	1	1	1
CGT	*2*	1	1	1	1
CIM	*2*	1	1	1	1
CLT	*2*	1	1	1	1
CVT	*2*	1	1	0	1
CAB	*2*	1	0	1	0

Table 1: Features (and their numerical values and weights) for subjective estimation for four close languages

Table-1 shows the features, the weights assigned to them, and their values for the four languages. Table-2 gives the weighted values of the features. As can be seen from this table, the relative costs of adapting resources of Telugu for Kannada, Malayalam and Tamil are 9, 20 and 19 respectively. These result seem to be intuitively correct.

	KN	ML	TM
DAG	3	3	3
DIM	0	6	0
FWO	0	0	4
DWS	0	9	6
CNT	0	0	0
CAT	0	0	0
CDT	2	2	2
CST	2	0	0
CGT	0	0	0
CIM	0	0	0
CLT	0	0	0
CVT	0	0	2
CAB	2	0	2
Total	9	20	19

Table 2: Differences of weighted feature values for Kannada, Malayalam and Tamil from Telugu

References

Robert Caldwell. 1913. *A Comparative Grammar of the Dravidian or South-Indian Family of Languages*. Kegan Paul, Trench, Trubner & and Co. Ltd., London.

I. Dyen, J.B. Kruskal, and P. Black. 1992. An indo-european classification: A lexicostatistical experiment. In *Transactions of the American Philosophical Society, 82:1-132*.

T. Mark Ellison and Simon Kirby. 2006. Measuring language divergence by intra-lexical comparison. In *Proceedings of ACL*, Sydney, Australia. Association for Computational Linguistics.

M. B. Emeneau. 1956. India as a linguistic area. In *Linguistics 32:3-16*.

April McMahon and Robert McMahon. 2005. *Language Classification by the Numbers*. Oxford University Press, Oxford.

J. Nerbonne. 2005. Review of ‘language classification by the numbers’ by april mcmahon and robert mcmahon.

Anil Kumar Singh and Harshit Surana. 2007. Can corpus based measures be used for comparative study of languages? In *Proceedings of the ACL Workshop Computing and Historical Phonology*, Prague, Czech Republic.

Anil Kumar Singh. 2006. A computational phonetic model for indian language scripts. In *Constraints on Spelling Changes: Fifth International Workshop on Writing Systems*, Nijmegen, The Netherlands.