Natural Language Processing for Less Privileged Languages: Where do we come from? Where are we going?

Anil Kumar Singh
Language Technologies Research Centre
IIIT, Hyderabad, India
anil@research.iiit.ac.in

Abstract

In the context of the IJCNLP workshop on Natural Language Processing (NLP) for Less Privileged Languages, we discuss the obstacles to research on such languages. We also briefly discuss the ways to make progress in removing these obstacles. We mention some previous work and comment on the papers selected for the workshop.

1 Introduction

While computing has become ubiquitous in the developed regions, its spread in other areas such as Asia is more recent. However, despite the fact that Asia is a dense area in terms of linguistic diversity (or perhaps because of it), many Asian languages are inadequately supported on computers. Even basic NLP tools are not available for these languages. This also has a social cost.

NLP or Computational Linguistics (CL) based technologies are now becoming important and future intelligent systems will use more of these techniques. Most of NLP/CL tools and technologies are tailored for English or European languages. Recently, there has been a rapid growth of IT industry in many Asian countries. This is now the perfect time to reduce the linguistic, computational and computational linguistics gap between the 'more privileged' and 'less privileged' languages.

The IJCNLP workshop on NLP for Less Privileged Language is aimed at bridging this gap. Only when a basic infrastructure for supporting regional languages becomes available can we hope for a more equitable availability of opportunities made possible by the language technology. There have already been attempts in this direction and this workshop will hopefully take them further.

Figure-1 shows one possible view of the computational infrastructure needed for language processing for a particular language, or more preferably, for a set of related languages.

In this paper, we will first discuss various aspects of the problem. We will then look back at the work already done. After that, we will present some suggestion for future work. But we will begin by addressing a minor issue: the terminology.

2 Terminology

There can be a debate about the correct term for the languages on which this workshop focuses. There are at least four candidates: less studied (LS) languages, resource scarce (RS) languages, less computerized (LC) languages, and less privileged (LP) languages. Out of these, two (LS and RS) are too narrow for our purposes. LC is admittedly more objective, but it also is somewhat narrow in the sense that it does not cover the lack of resources for creating resources (finance) and the lack of linguistic study. We have used LP because it is more general and covers all the aspects of the problem. However, it might be preferable to use LC in many contexts.

As the common element among all these terms is the adjective 'less' ('resoure scarce' can be paraphrased as 'with less resources'), perhaps we can avoid the terminological debate by calling the languages covered by any such terms as the L-languages.

Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 7–12,
Hyderabad, India, January 2008.

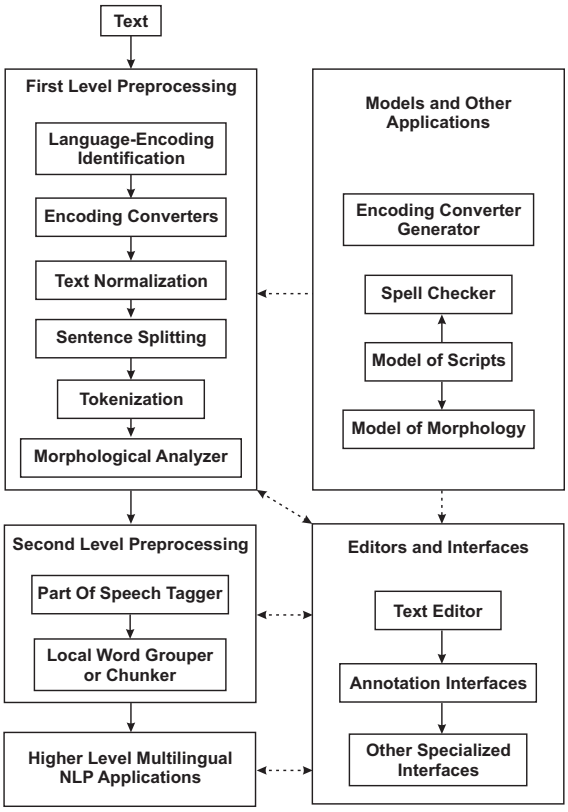

Figure 1: One view of the basic computational infrastructure required for Natural Language Processing or Computational Linguistics. Components like encoding converters are needed for languages with less standardization, such as the South Asian languages. Language resources like lexicon, corpora etc. have not been shown in this figure.

3 Problems

Not surprisingly, the terms mentioned in the previous section cover different aspects of the problems that restrict work on and for these languages. There is a lack of something and each of those terms covers some part of what is lacking.

3.1 Linguistic Study

The term LS languages indicates that these are not well studied linguistically. The sheer amount of linguistic analysis available for English is so huge that the linguistic work on even a language like Hindi, which is spoken or understood by a billion people, is simply not comparable. For languages (or dialects) like Santali or Manipuri, the situation is much worse. And there are a large number of languages which have been studied even less than Santali or Manipuri. There are dozens (more accurately, hundreds) of such languages in South Asia alone[1]. It can be said that very little is known about the majority of languages of the world, many of which are facing extinction.

3.2 Language Resources

Even those languages which have been studied to a good extent, e.g. Telugu, lack language resources, e.g. a large dictionary in machine readable form, let alone resources like WordNet or FrameNet, although efforts are being made to develop resources for some of these languages. The term RS covers this aspect of the problem.

3.3 Computerization

Computerization, in general, might include machine readable language resources and NLP tools etc., but here we will restrict the meaning of this term to the support for languages that is provided on computers, either as part of operating systems, or in the commonly used applications such as word processors. In the narrowest sense, computerization means language-encoding support. Even this level of support is currently not available (or is inadequate) for a large number of languages.

3.4 Language Processing

Proper computerization (in the restricted sense) is a prerequisite to effective language processing. But even without adequate computerization, attempts are being made towards making language processing possible for the L-languages. However, language processing for the L-languages is still far behind that for English. For a large number of language it is, in fact, non-existent. This is true even for a language like Gujarati, which is the official language of the state of Gujarat in India and is recognized as a scheduled language by the government of India. And it is actually used as the first language by the people of Gujarat, which is one of the larger states in India. While adequate computerization may be easy to achieve in the near future, at least theoretically, language processing (and building language resources) is going to be much more difficult task.

[1] Ethnologue: http://www.ethnologue.com/web.asp

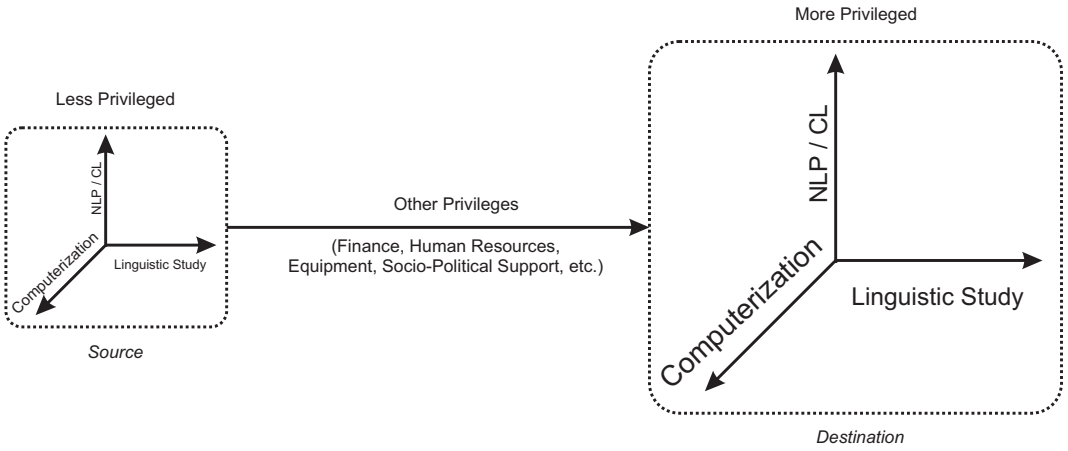

Figure 2: The four dimensions of the problem: The *Source* is where we come from and *Destination* is where we are going. The problem is to go from the *Source* to the *Destination* and the solution is non-trivial.

3.5 Other Privileges

One of the major reasons why building language resources and providing language processing capabilities for the L-languages is going to be a very difficult task is the fact that these languages lack the privileges which make it possible to build language resources and NLP/CL tools. By 'privileges' we mean the availability of finance, equipment, human resources, and even political and social support for reducing the lack of computing and language processing support for the L-languages. The lack of such 'privileges' may be the single biggest reason which is holding back the progress towards providing computing and language processing support for these languages.

4 Some (Partially) Successful Efforts

The problem seems to be insurmountable, but there has been some progress. More importantly, the urgency of solving this problem (even if partially) is being realized by more and more people. Some recent events or efforts which tried to address the problem and which have had some impact in improving the situation are:

- The LREC conferences and workshops[2].
- Workshop on "Shallow Parsing in South Asian Languages", IJCAI-07, India.
- EMELD and the Digital Tools Summit in Linguistics, 2006, USA.
- Workshop on Language Resources for European Minority Languages, 1998, Spain.
- Projects supported by ELRA on the Basic Language Resource Kit (BLARK) that targets the specifications of a minimal kits for each language to support NLP tools development[3].
- There is also a corresponding project at LDC (the Less Commonly Taught Languages[4]).
- The IJCNLP Workshop on Named Entity Recognition for South and South Asian Languages[5].

This list is, of course, not exhaustive. There are many papers relevant to the theme of this workshop at the IJCNLP 2008 main conference[6], as at some previous major conferences. There is also a very relevant tutorial (Mihalcea, 2008) at the IJCNLP 2008 conference about building resources and tools for languages with scarce resources.

Even the industry is realizing the importance of providing computing support for some of the L-languages. In the last few years there have been many announcements about the addition of some

[2] www.lrec-conf.org

[3] http://www.elda.org/blark

[4] http://projects.ldc.upenn.edu/LCTL

[5] http://ltrc.iiit.ac.in/ner-ssea-08/

[6] http://ijcnlp2008.org

such language to a product or a service and also of the addition of better facilities (input methods, transliteration, search) in an existing product or service for some L-language.

5 Towards a Solution

Since the problem is very much like the conservation of the Earth's environment, there is no easy solution. It is not even evident that a complete solution is possible. However, we can still try for the best possible solution. Such a solution should have some prerequisites. As Figure-2 shows, the 'other privileges' dimension of the problem has to be a major element of the solution, but it is not something over which researchers and developers have much control. This means that we will have to find ways to work even with very little of these 'other privileges'. This is the key point that we want to make in this paper because it implies that the methods that have been used for English (a language with almost unlimited 'privileges') may not be applicable for the L-languages. Many of these methods assume the availability of certain things which simply cannot be assumed for the L-languages. For example, there is no reasonable ground to assume that there will be (in the near future) corpus even with shallow levels of annotation for Avadhi or Dogri or Konkani, let alone a treebank like resource. Therefore, we have to look for methods which can work with unannotated corpus. Moreover, these methods should also not require a lot of work from trained linguists because such linguists may not be available to work on these languages. There is one approach, however, that can still allow us to build resources and tools for these languages. This is the approach of adapting the resources of a linguistically close but more privileged language. It is this area which needs to be studied and explored more thoroughly because it seems to be the only practical way to make the kind of progress that is required urgently. The process of resource adaptation will have to studied from linguistic, computational, and other practical points of view. Since 'other privileges' are a major factor as discussed earlier, some ways of calculating the cost of adaptation have also to be found.

Another very general but important point is that we will have to build multilingual systems as far as possible so that the cost per language is reduced. This will require innovation in terms of modeling as well as engineering.

6 Some Comments about the Workshop

The scope of the workshop included topics such as the following:

- Archiving and creation of interoperable data and metadata for less privileged languages
- Support for less privileged language on computers. This includes input methods, display, fonts, encoding converters, spell checkers, more linguistically aware text editors etc.
- Basic NLP tools such as sentence marker, tokenizer, morphological analyzer, transliteration tools, language and encoding identifiers etc.
- Advanced NLP tools such as POS taggers, local word grouper, approximate string search, tools for developing language resources.

There were a relatively large number of submissions to the workshop and the overall quality was at least above average. The most noteworthy fact is that the variety of papers submitted (and selected) was pleasantly surprising. The workshop includes paper on topics as diverse as Machine Translation (MT) from text to sign language (an L-language on which very few people have worked) to MT from speech to speech. And from segmentation and stemming to parser adaptation. Also, from input methods, text editor and interfaces to part of speech (POS) tagger. The variety is also remarkable in terms of the languages covered and research locations.

In addition, the workshop includes three invited talks: the first on building language resources by resource adaptation (David and Maxwell, 2008); the second on cross-language resource sharing (Sornlertlamvanich, 2008b); and the third on breaking the Zipfian barrier in NLP (Choudhury, 2008). It can be said that the workshop has been a moderate success. We hope it will stimulate further work in this direction.

7 An Overview of the Papers

We noted above that resource adaptation needs a lot more study. In one of the papers at the workshop, Zeman and Resnik presented their work on cross-language parser adaptation between related languages, which can be highly relevant for the L-languages in 'linguistic areas' (Emeneau, 1956; Emeneau, 1980). Maxwell and David suggest a better way to weave together a descriptive grammar with a formal grammar through collaboration between linguists and computer scientists. Alegria et al. discuss the strategies for sustainable MT for Basque. They suggest that the main elements of such a strategy should be incremental design, reusability, standardization and open source development.

Among the papers which focus more on computerization and building of tools, Sornlertlamvanich et al. present a ubiquitous system called KUI for collective intelligence development. Goonetilleke et al. describe a predictive text input system called SriShell Primo for Sinhala language. Veeraraghavan and Roy describe a text editor and a framework for working with Indic scripts. Aggarwal and Dave present an implementation of a speech recognition system interface for Indian languages.

Riza presents brief overview of the literature on language endangerment, with focus on the Indonesian languages. Some other papers focused more on linguistic study as applied for computational purposes. Among them, Ali et al. investigate the optimal order of factors for the computational treatment of personal anaphoric devices in Urdu discourse. Muhirwe and Trosterud discuss finite state solutions for reduplication in Kinyarwanda language. Maung Maung and Mikami describe a rule-based syllable segmentation of Myanmar text. In another paper on a related domain, Sarkar and Bandyopadhyay present a design of a rule-based stemmer for natural language text in Bengali.

Among the papers focusing more on NLP, Dasgupta et al. present a prototype machine translation system from text to Indian Sign Language (ISL). In another paper on MT, Ellis et al. describe an Finnish to English speech to speech machine translation system that they have currently tried with some success on the Bible. Doren and Bandyopadhyay present a morphology driven Manipuri POS tagger. Another paper on POS tagging is by Patel and Gali. They have tried to build a tagger for Gujarati.

8 Conclusion

We discussed the problem of the lack of linguistic study, language resources, NLP tools for some languages, which we called the L-languages since they **l**ess of something. We argued that the 'other privileges' form another dimension of the problem and are a crucial factor in deciding what methods we should use to solve this problem. The technical has to take into account this non-technical factor. We suggested that resource adaptation may be one to move forward. Finally we made some comments about the NLPLPL-08 workshop.

9 Acknowledgment

We would specially like to thank Samar Husain and Harshit Surana (Language Technologies Research Centre, IIIT, Hyderabad, India) for providing vital help in organizing this workshop.

References

Rajesh Kumar Aggarwal and Mayank Dave. 2008. Implementing a speech recognition system interface for indian languages. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

I Alegria, Xabier Arregi, Xabier Artola, Arantza Diaz de Ilarraza, Gorka Labaka, Mikel Lersundi, Aingeru Mayor, and Kepa Sarasola. 2008. Strategies for sustainable mt for basque: incremental design, reusability, standardization and open-source. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

Mohammad Naveed Ali, Muhammad Abid Khan, and Muhammad Aamir Khan. 2008. An optimal order of factors for the computational treatment of personal anaphoric devices in urdu discourse. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

Monojit Choudhury. 2008. Breaking the zipfian barrier of nlp. *Invited Talk at the IJCNLP Workshop on NLP for Less Privileged Languages*. Hyderabad, India.

Tirthankar Dasgupta, Sandipan Dandapat, and Anupam Basu. 2008. Prototype machine translation system from text-to-indian sign language. In *Proceedings*

of the IJCNLP Workshop on NLP for Less Privileged Languages, Hyderabad, India.

Anne David and Michael Maxwell. 2008. Building language resources: Ways to move forward. *Invited Talk at the IJCNLP Workshop on NLP for Less Privileged Languages, 2008*. Hyderabad, India.

Timo Honkela David Ellis, Mathias Creutz and Mikko Kurimo. 2008. Speech to speech machine translation: Biblical chatter from finnish to english. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

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

M. B. Emeneau. 1980. *Language and linguistic area. Essays by Murray B. Emeneau. Selected and introduced by Anwar S. Dil*. Stanford University Press.

Sandeva Goonetilleke, Yoshihiko Hayashi, Yuichi Itoh, and Fumio Kishino. 2008. Srishell primo: A predictive sinhala text input system. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

Zin Maung Maung and Yoshiki Mikami. 2008. A rule-based syllable segmentation of myanmar text. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

Michael Maxwell and Anne David. 2008. Joint grammar development by linguists and computer scientists. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

Rada Mihalcea. 2008. How to add a new language on the nlp map: Building resources and tools for languages with scarce resources. *Tutorial at the Third International Joint Conference on Natural Language Processing (IJCNLP)*. Hyderabad, India.

Jackson Muhirwe and Trond Trosterud. 2008. Finite state solutions for reduplication in kinyarwanda language. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

Chirag Patel and Karthik Gali. 2008. Part of speech tagger for gujarati using conditional random fields. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

Hammam Riza. 2008. Indigenous languages of indonesia: Creating language resources for language preservation. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

Sandipan Sarkar and Sivaji Bandyopadhyay. 2008. Design of a rule-based stemmer for natural language text in bengali. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2008. Morphology driven manipuri pos tagger. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

Virach Sornlertlamvanich, Thatsanee Charoenporn, Kergrit Robkop, and Hitoshi Isahara. 2008a. Kui: an ubiquitous tool for collective intelligence development. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

Virach Sornlertlamvanich. 2008b. Cross language resource sharing. *Invited Talk at the IJCNLP Workshop on NLP for Less Privileged Languages, 2008*. Hyderabad, India.

Krishnakumar Veeraraghavan and Indrani Roy. 2008. Acharya - a text editor and framework for working with indic scripts. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*, Hyderabad, India.