On the Role of NLP in Linguistics

Dipti Misra Sharma
Language Technologies Research Centre
IIIT-H, Hyderabad, India
dipti@iiit.ac.in

Abstract

This paper summarizes some of the applications of NLP techniques in various linguistic sub-fields, and presents a few examples that call for a deeper engagement between the two fields.

1 Introduction

The recent success of data-driven approaches in NLP has raised important questions as to what role linguistics must now seek to play in further advancing the field. Perhaps, it is also time to pose the same question from the other direction: As to how NLP techniques can help linguists make informed decisions? And how can the advances made in one field be applied to the other?

Although, there has been some work on incorporating NLP techniques for linguistic fieldwork and language documentation (Bird, 2009), the wider use of NLP in linguistic studies is still fairly limited. However, it is possible to deepen the engagement between the two fields in a number of possible areas (as we shall see in the following sections), and gain new insights even during the formulation of linguistic theories and frameworks.

2 Historical Linguistics and Linguistic Typology

Computational techniques have been successfully used to classify languages and to generate phylogenetic trees. This has been tried not just with handcrafted word lists (Atkinson et al., 2005; Atkinson and Gray, 2006; Huelsenbeck et al., 2001) or syntactic data (Barbaçon et al., 2007) but with lists extracted from written corpus with comparable results (Rama and Singh, 2009; Singh and Surana, 2007). These techniques are inspired from the work in computational phylogenetics, which was aimed at constructing evolutionary trees of biological species. Constructing a phylogenetic tree for languages usually requires the calculation of distances between pairs of languages (usually based on word lists). These distances are then given as input to a computational phylogenetic algorithm. Their successful use for languages has opened the possibility of using computational techniques for studying historical linguistics. They have already been used for estimating divergence times of language families (Atkinson et al., 2005). Figure 1 shows a phylogenetic tree created using feature n-grams (Rama and Singh, 2009).

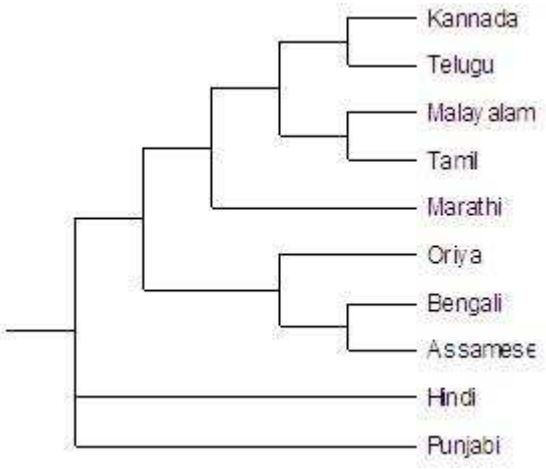

Figure 1: Phylogenetic tree using feature n-grams

Another area for the application of NLP techniques is language typology. For example, linguistic similarity and its estimation can be seen as fundamental ideas in NLP. The systematic study of different kinds of linguistic similarity offers insights towards the theoretical studies of languages (Singh, 2010). In brief, the typology of linguistic similarity for computational purposes is related to linguistic levels (depth), differences among languages (linguality) and linguistic units (granularity). Thus, language can be seen as a system of symbols whose meanings are defined

Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, ACL 2010, pages 18–21,
Uppsala, Sweden, 16 July 2010.

in terms of their estimated similarity and distance with other symbols. Can this, together with what Cognitive Linguists have been studying (Robinson and Ellis, 2008), which also involves linguistic similarity, often directly, have some relevance for linguists?

3 Lexical Correspondence and Linguistic Units

A further case in point is lexical correspondence across languages, which poses a problem for cross-lingual and multilingual applications. To address this and some other issues, a linguistic *unit* that behaves similarly across languages can be conceptualized. Such a unit, may include morphological variation (inflectional and derivational), compounds, multi word expressions etc. as in the Hindi and Telugu examples below:

- Single token content words: *raama, raama* (Ram); *vah, atanu* (he); *vyakti, manishii* (person) etc.
- Nouns with inflections: *bacce, pillalu* (children); *bacce ko, pillalaki* (to the child); *raama se, raamudunundii* (from Rama) etc.
- Verbs with inflections and tense, aspect and modality (TAM) markers: *karnaa-caahiye, cayiyaalii* (should do); *ho sakataa thaa, ayyiyedemo* (could have happened) etc.
- Multi word expressions such as idioms, phrasal verbs and 'frozen expressions': *pahaaD toDanaa* (breaking mountains); *muNha ki khaana* (getting defeated) etc.
- Compounds: *jaati-prathaa* (caste system); *vesh-bhuushaaoN* (dresses); *akkaDaa-ikkaDaa* (here and there) etc.

This unit might, among other things, form the basis of the structure of lexical resources, such that these resources have a direct correspondence across languages. This can further facilitate comparative study of languages (Singh, 2010).

4 Applications

Computational techniques can also be used to design tools and material for language learning and teaching. Here games can play a useful role. Although, a large number of online games are available, most of them do not use the latest language processing techniques. Games can also be used to generate language resources.

The core idea in Human Computation (Von Ahn, 2005) is that computers should do what they do best and that humans seamlessly work with them to do what computers cannot. One of the ways to merge the two is in the form of carefully designed games.

Another insight comes from Machine Translation. More than any other sub-field in NLP, it is the data-driven approaches to machine translation that have proven to be particularly successful over the past few years. We have been exploring various approaches towards hybridization of our rule-based MT system. Building the transfer-grammar of such systems is perhaps one of the most time-intensive tasks that involves careful analysis of test data. However, data driven techniques can come to the aid of linguists in this case. The recent work on automatic acquisition of rules from parallel corpora (Lavie et al., 2004) can help identify a large number of common syntactic transformations across a pair of languages, and help unearth those transformations that might otherwise be missed by a rule-based grammar. They can be further used to prioritize the application of rules based on the observed frequencies of certain syntactic transformations.

5 NLP Tools and Linguistics

NLP techniques draw features from annotated corpora which are a rich linguistic resource. However, these corpora can also be used to extract grammars, which on one hand feed the parser with features (Xia, 2001), and on the other, act as a resource for linguistic studies. For example, in Hindi dependency parsing the use of vibhakti (post-positions) and TAM labels has proven to be particularly useful even in the absence of large amounts of annotated corpora (Ambati et al., 2010). This also helped bring to light those features of the grammar that govern certain structure choices and brought to notice some previously overlooked linguistic constructions. Thus, the result is an iterative process, where both the grammar and the features are refined.

Discourse Processing is another rapidly emerging research area with considerable potential for interaction and collaboration between NLP and Linguistics. In the absence of fully developed theories/frameworks on both sides, focus on syner-

gizing research efforts in the two disciplines (such as devising novel ways to empirically test linguistic hypotheses) from the initial stage itself, can yield a substantially richer account of Discourse.

Linguistic theories are formalized based on observations and abstractions of existing linguistic facts. These theories are then applied to various languages to test their validity. However, languages throw up new problems and issues before theoreticians. Hence, there are always certain phenomena in languages which remain a point of discussion since satisfactory solutions are not available. The facts of a language are accounted for by applying various techniques and methods that are offered by a linguistic framework. For example, syntactic diagnostics have been a fairly reliable method of identifying/classifying construction types in languages. They work fairly well for most cases. But in some cases even these tests fail to classify certain elements. For example, Indian languages show a highly productive use of complex predicates (Butt, 1995; Butt, 2003). However, till date there are no satisfactory methods to decide when a noun verb sequence is a 'complex predicate' and when a 'verb argument' case. To quote an example from our experience while developing a Hindi Tree Bank, annotators had to be provided with guidelines to mark a N V sequence as a complex predicate based on some linguistic tests. However, there are instances when the native speaker/annotator is quite confident of a construction being a complex predicate, even though most syntactic tests might not apply to it.

Although, various theories provide frames to classify linguistic patterns/items but none of them enables us to (at least to my knowledge) handle 'transient/graded' or rather 'evolving' elements. So, as of now it looks like quite an arbitrary/adhoc approach whether to classify something as a complex predicate or not. In the above cited example, the decision is left to the annotator's intuition, since linguists don't agree on the classfication of these elements or on a set of uniform tests either. Can the insights gained from inter-annotator agreement further help *theory* refine the diagnostics used in these cases? And can NLP techniques or advanced NLP tools come to the aid of linguists here? Perhaps in the form of tools that can (to an extent) help automate the application of syntactic diagnostics over large corpora?

6 Collaborations

Interdisciplinary areas such as Computational Linguistics/NLP need a much broader collaboration between linguists and computer scientists. Experts working within their respective fields tend to be deeply grounded in their approaches towards particular problems. Also, they tend to speak different 'languages'. Therefore, it becomes imperative that efforts be made to bridge the gaps in communication between the two disciplines. This problem is all the more acute in India, since the separation of disciplines happens at a very early stage. Objectives, goals, methods and training are so different that starting a communication line proves to be very difficult. Thus, it is important for those people who have synthesised the knowledge of the two disciplines to a large degree, to take the lead and help establish the initial communication channels. Our own experiences while devising common tagsets for Indian languages, made us realize the need for both linguistic and computational perspectives towards such problems. While a linguist's instinct is to look for exceptions in the grammar (or any formalism), a computer scientist tends to look for rules that can be abstracted away and modeled. However, at the end, both ways of looking at data help us make informed decisions.

Acknowledgements

Many thanks to Dr. Rajeev Sangal, Anil Kumar Singh, Arafat Ahsan, Bharath Ambati, Rafiya Begum, Samar Husain and Sudheer Kolachina for the discussions and inputs.

References

B.R. Ambati, S. Husain, J. Nivre, and R. Sangal. 2010. On the role of morphosyntactic features in Hindi dependency parsing. In *The First Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, page 94.

QD Atkinson and RD Gray. 2006. How old is the Indo-European language family? Progress or more moths to the flame. *Phylogenetic Methods and the Prehistory of Languages (Forster P, Renfrew C, eds)*, pages 91–109.

Q. Atkinson, G. Nicholls, D. Welch, and R. Gray. 2005. From words to dates: water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society*, 103(2):193–219.

S. Bird. 2009. Natural language processing and linguistic fieldwork. *Computational Linguistics*, 35(3):469–474.

M. Butt. 1995. *The structure of complex predicates in Urdu*. Center for the Study of Language and Information.

M. Butt. 2003. The light verb jungle. In *Workshop on Multi-Verb Constructions*. Citeseer.

J.P. Huelsenbeck, F. Ronquist, R. Nielsen, and J.P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550):2310–2314.

A. Lavie, K. Probst, E. Peterson, S. Vogel, L. Levin, A. Font-Llitjos, and J. Carbonell. 2004. A trainable transfer-based machine translation approach for languages with limited resources. In *Proceedings of Workshop of the European Association for Machine Translation*. Citeseer.

Taraka Rama and Anil Kumar Singh. 2009. From bag of languages to family trees from noisy corpus. In *Proceedings of the Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

Peter Robinson and Nick Ellis. 2008. *Handbook of Cognitive Linguistics and Second Language Acquisition*. Routledge, New York and London.

Anil Kumar Singh and Harshit Surana. 2007. Can corpus based measures be used for comparative study of languages? In *Proceedings of the Ninth Meeting of ACL Special Interest Group on Computational Phonology and Morphology*, Prague, Czech Republic. Association for Computational Linguistics.

Anil Kumar Singh. 2010. *Modeling and Application of Linguistic Similarity*. Ph.D. thesis, IIIT, Hyderabad, India.

Luis Von Ahn. 2005. *Human computation*. Ph.D. thesis, Pittsburgh, PA, USA. Adviser-Blum, Manuel.

Fei Xia. 2001. *Automatic Grammar Generation from Two Different Perspectives*. Ph.D. thesis, University of Pennsylvania.