Aspects of Language Resource Management: Creation and Utilisation

Tokunaga, Takenobu – Department of Computer Science,
Tokyo Institute of Technology, Japan
take@cl.cs.titech.ac.jp

I would like to raised issues concerning language resource management, from two viewpoints: (1) creating language resources, and (2) utilising language resources.

Language Resource Creation

Language resource creation involves various kinds of concrete and abstract entities, spanning from project managers and annotators, to documents and annotation schemata/guidelines used in annotating them. The relations between these entities are also far from trivial. For instance, the creation of a language resource might involve the use of several annotation schemata (defined as tagsets); an annotator might be assigned to work on different annotation tasks (using different/multiple tagsets); and so likewise a document might be annotated with different/multiple annotation schemata; it is also possible that annotations might span over multiple documents, and so on.

It is no longer an uncommon practice to create a new language resource on top of an existing language resource by adding an additional layer of annotation suited to a new task, e.g. adding named entity annotations upon an already morpho-syntactically annotated corpus. Such layered annotation further complicates the relationships between the schematic entities, and also between the schematic entities and real-world ones that manage them. Proper management of these entities and their relations can, therefore, substantially contribute to keeping the resultant language resources consistent, and improve their overall quality. As an example, creating schematic constraints that named entity tags can be annotated only on elements with certain types of pre-existing annotations (e.g. *np* tags) would prevent many careless annotator mistakes. Despite this, the importance of proper management of language resource creation has attracted less attention and been largely overlooked in our community. A standard or at least a reference model for the management of the language resource creation process should be considered.

Language Resource Utilisation

Corpus-based approaches have been a main stream of Human Language Technology (HLT) for the last two decades. These approaches always involve language resources and machine learning techniques to utilise them. Having relevant language resources has, therefore, become just as crucial as inventing a new method to achieve a given research goal. In some cases, the preparation of relevant language resources alone, when used with an off-the-shelf machine learning technique, may be sufficient for the task. In this context, consolidating an environment in which researchers can easily find relevant language resources for their purpose is a must. Currently many attempts have been made for standardising metadata of language resources and cataloging them. Metadata currently under discussion, however, may be insufficient to meet the demands of research. Looking into the future, we need to realise a query mechanism where relevant language resources can be retrieved by queries specified in natural language, such as "I would like to have a dialogue corpus in which such and such information is annotated.".

A related but different issue is making language resources citable. As described above, language resources increasingly account for a large amount of importance in research, and thus also in research papers, which describe the results. Unfortunately we have no standard way of citing them. As a result, in many cases language resources are cited by indicating the URLs of their distribution sites or by referring to the research papers describing them. It is high time we make language resources citable first-class objects.