Intelligence Resource Collection for Low-Density Languages

Jaime Carbonell
Language Technologies Institute, Carnegie Mellon University

Whereas there are over 6,000 languages[1], less than 100 have ample written resources that are easily collected. The next tranche of several hundred languages may have significant monolingual resources, but very little parallel text that could be used to train corpus-based machine translation. Beyond the top 1,000 languages even monolingual text is rather scarce. First, we address emerging language technologies, in particular MT methods, that require fewer scarce resources, focusing on emerging new paradigms for MT based on richer linguistic priors, and consequent collection of human judgments primarily from bilinguals (e.g. translations, alignments, corrections, or heavier linguistic annotation) to supplement any small-scale existing collections that may exist for the languages of interest. Then, we address new methods collect and augment linguistic resources dynamically via active and proactive machine learning, so that the linguistic information may be elicited from large distributed networks of informants.

Whereas the dominant paradigm for machine translation remains phrase-based statistical MT, some interesting variants have emerged, including syntax-augmented statistical SMT[2], where the statistical translation model is trained over grammatical as well as lexical structures, yielding better MT performance for some distance language pairs such as Chinese-English and Urdu-English. Another corpus-based method, context-based MT[3], requires a thorough bilingual dictionary and a large target corpus, but no parallel text, thus is well suited for translating from a low-resource into a high-resource language. New research is focusing on combining linguistic and statistical methods to generate effective MT for lower-density languages, essentially incorporating linguistic generalizations in its translation and language models. All these approaches signal evolving needs for resource creation and collection: Rather than ever-growing bilingual text with diminishing accuracy gains in phrasal SMT, tree-banks to train parsers and linguistic models, and comprehensive bilingual dictionaries might prove more useful in order to support ongoing research and yield new results in language technologies.

The advent of crowd-sourcing such as Amazon's Mechanical Turk, and data-collection games-with-a-purpose[4], suggests new highly-scalable cost-effective ways of collecting linguistic information. However, the reliability and utility of linguistic data so gathered is highly suspect, as un-vetted individuals may lack requisite expertise, or worse yet, as crowds may attempt to subvert ("game") the system, for instance by cutting and basting online MT system output to provide low-quality translations quickly, instead of providing high-quality translation required

[1] http://www.ethnologue.com/ethno_docs/distribution.asp?by=size

[2] Knight "Capturing practical natural language transformations" *Machine Translation*, 12(2), 2007. Also Zollmann *et al.* "Syntax-Augmented MT" ACL Workshop on SMT, 2007.

[3] Carbonell et al, "Context-based machine translation" Proc of AMTA, 2006.

[4] Von Ahn, "Games with a purpose", *Invisible Computing,* 2006.

to build or extend a parallel training corpus. However, some relief is offered by the emerging field of proactive machine learning[5], which attempts to learn annotator accuracy and consistency, and directs annotation requests to the most appropriate and cost-effective information sources dynamically. Essentially the annotation task is cast as an information-theoretic optimization problem with partial (but increasing) knowledge of external information sources. This new method has implications for building linguistic resources with large networks of distributed annotators and information providers – not just via fleeting encounters with information providers such as in AMT or web-games, but perhaps more importantly for establishing distributed networks of annotators with emerging and/or differential expertise and with different cost-benefit structures.

[5] Donmez & Carbonell, "Proactive learning: Cons-sensitive active learning with multiple imperfect oracles. Proc. of CIKM, 2008. Also Baldridge & Palmer, "How well does active learning actually work?" Proc of EMNLP, 2009.