Language technology from scratch

Kevin P. Scannell

May 11, 2008

This is the transcript of a video lecture shown at the Vox Humanitatis conference on technology for minority languages, held as part of the 41st Festival of the Piedmont Region, Cherasco, Italy.

Hi, this is Kevin Scannell, I'm coming to you from St. Louis, Missouri in the United States where I am Director of Computer Science at Saint Louis University. First of all I am grateful to Vox Humanitatis for organising this wonderful conference and for giving me the opportunity to talk about my software. I have been developing language technology for minority languages and underresourced languages for almost ten years now. Initially for Irish and the other Celtic languages and in recent years on a much broader scale. I want to start out by sharing some of the underlying philosophy of this work because I believe these philosophies are in a sense just as important as the tools that I have developed.

First and foremost I release everything I do as free software. A language belongs to the people who speak it and so should the linguistic data needed for developing language technology. And in practical terms sharing tools and data is the only realistic way forward given the scarce resources that most groups face. Open Source avoids reduplication of effort and saves everyone time and money. Plus, it makes for better software.

I also have a web crawler that is gathering texts right now for 419 different languages. These texts can be used in a number of different ways to develop spell checkers, grammar checkers and more advanced tools. Of special interest in terms of statistical methods are unsupervised machine learning algorithms which don't require manually annotated data to work. Which is good since such data don't exist for underresourced languages by definition. It's also possible to carry over advanced technologies like parsers or semantic networks from global languages like English, French, Chinese etc. to underresourced languages given enough parallel text between the two languages. I'll give an example of this later. Finally, we are all facing the same obstacles of bringing language technology to our languages and I believe an effective strategy is for many different language groups to co-operate and collaborate on the development of language-independent tools that language-specific data can be fed into. The standard open source spellchecking engines Aspell and Hunspell are like this, as is my web crawler and my grammar checking engine An Gramadóir.

An Gramadóir is a framework that allows you to develop language technology from scratch, which is a good thing, because in fact most groups start with absolutely nothing, maybe just a small list of words. We start this process by encoding morphological rules, prefixes and suffixes and the like, in a format which is understood by many open source tools. Scripts combine these rules with statistics from my web corpora to find candidate "root words" which are validated by hand by native speakers. Unless the language has unusually complex morphology this process produces reliable spellcheckers with a minimum of effort. The proof is in the pudding; we've created 19 new open source spell checkers and most of these since 2004. Once a spellchecker is in place, we move on to part-of-speech tagging which is a central part of the grammar checking process. First we assign all possible part of speech tags to all words using the morphological rules plus some hand tagging for closed class words like prepositions, etc. Then these tags, plus the web-crawled corpora from before are fed into an unsupervised machine learning algorithm to produce a reliable part of speech tagger automatically.

An Gramadóir has an easy to use web interface. Just type or copy/paste the text you want to grammar check into the box. Select the language you'd like the error messages to appear in and click "send". After a second a page appears with the errors highlighted in red together with short explanations in the language you chose.

I'd like to close with a more advanced resource developed within this framework, namely, a semantic network which is simply a database of words and semantic relationships between them. For example synonyms and antonyms are marked as they would be in a standard thesaurus. But semantic networks usually also mark richer relationships like hypernyms and hyponyms, these are broader and narrower terms, or meronyms and holonyms, part vs. whole relationships. Semantic networks have many important applications in natural language processing. I've created a network for Irish using unsupervised learning techniques from the existing English WordNet and a large Irish and English parallel corpus. This is a good example of the philosophy mentioned earlier: exploit resources from global languages for your own ends.

The best part of all is that you can browse the WordNet using an open source 3D browser called “Morcego”. Put your word in the search box and you will see a graph with red and green nodes. The red nodes represent orthographic words and the green nodes represent core senses in the language. It is possible to get a better view of the graph simply by dragging your mouse and rotating. In this case the orthographic word “bonn” also represents the core sense “sole (of the foot)”; it also means “track” or “footprint”, or “medal, medallion”. Clicking a node centers it and shows a number of other nodes with related meanings. In this case “(an) honor, distinction”. This is a powerful way to explore the semantic richness of a language. And especially so for endangered languages, when this richness can slip away in just a generation or two, which has certainly happened in the case of Irish.

Well, that’s all I have so thanks for listening and I hope you’ll get in touch with me if you are interested in using these tools or helping to develop them. Another thanks to Vox Humanitatis for this invitation and a big thanks to the many people all around the world who have collaborated with me on these tools over the past five years, it’s been great fun. Slán!