Automatic Lexical Acquisition - Bridging Research and Practice

Anna Korhonen
University of Cambridge
Computer Laboratory and RCEAL

There is a pressing need to develop comprehensive and accurate lexical resources for natural language systems dealing with real-world applications (e.g. high quality lexicons for information extraction and machine translation). Such resources are critical for enhancing the performance of systems and for improving their portability between domains. Currently, most lexical resources are developed manually by linguists. Manual work is costly, and the resulting resources require extensive labour-intensive porting to new tasks. Automatic acquisition or updating of lexical information from repositories of text (e.g. corpora, the web) is a more cost-effective approach to take. The approach is now increasingly viable given recent advances in Natural Language Processing (NLP) and machine learning technology. Yet, despite two decades of intensive research effort, hardly any acquisition technology has moved from research laboratories into widespread application. This holds back the development of language technology and makes it increasingly difficult to obtain funding for the research area. We can tackle this situation by concentrating research in viable areas which can benefit real world applications and act as a proof of concept for the line of research:

Focus of lexical acquisition

To date, research has been conducted in various areas of lexical acquisition, ranging from shallow to deep (e.g. terms, collocations, subcategorization frames, lexical-semantic classes, diathesis alternations, predicate-argument structures, word senses). While considerable research effort is required to improve performance in most areas of lexical acquisition, it is important to concentrate effort on the types of lexical information which can be acquired from large data sets with promising accuracy and which we know can benefit real-world applications the most.

Acquisition techniques

One of the biggest current research challenges is to improve the accuracy of existing techniques further and to replace small-scale techniques with more powerful and portable techniques. Without this leap, the technologies will always be limited in what they can achieve. For example,

- Instead of focusing on one type of lexical information (e.g. syntactic), we could integrate the acquisition of different types of lexical information (e.g. syntactic and semantic) so that they can support each other.

- Instead of conducting incremental research using existing methodology which we know will not transform the field, we could actively search for better suited and developed methodology in neighbouring fields where much of the existing methodology originally comes from (e.g. machine learning, engineering, physics).

- Instead of hoping that quantity equals quality, we could investigate the optimal balance between the two, and where quantity exceeds quality, develop sophisticated filtering techniques.

- Instead of developing approaches for supervised domain adaptation, we could focus on more realistic domain-adaptation which can deal with a small amount of training data. We should also investigate the minimum effort required to obtain the training data from users, the web, etc.

Multi-lingual acquisition

Much of the currently available technology has been evaluated for major languages (or for English) only. Evaluating the applicability of the techniques to other languages would be critical for both theoretical and practical reasons; for 1) improving the accuracy, scalability and robustness of the techniques, 2) advancing work in other languages, 3) gaining a better understanding of the language-specific / cross-linguistic components of lexical information, and 4) improving the performance of (multilingual) NLP applications (e.g. MT, IE).

Evaluation and real-world application

Many techniques are evaluated against the same lexical resources which (being incomplete, inaccurate, unsuitable for domains, and lacking frequency information) have motivated the very development of lexical acquisition. There is a need to investigate how to obtain more accurate evaluation data with the aid of users, experts, and automatic methods. Also, although automatically acquired lexical (frequency) information is potentially useful for many applications, its practical usefulness remains largely undemonstrated. It would be critical to conduct evaluation in the context of real-world tasks (e.g. information extraction, machine translation, text classification) on general and domain data, and within and across languages.

Recent research shows that even when not fully accurate, automatically acquired lexical information can be useful. It is therefore important to move beyond experimental research and use the most promising of the current technology to acquire lexical resources where they are needed the most in both research and development. Equally important is to use the techniques to obtain (statistical) information for improving and tuning existing manually built lexical resources for different tasks. For maximum impact, we should make the techniques and resources developed available for wider academic and industrial communities and encourage their use e.g. via the internet.