Research Exchange Program (ReX) Update from the Field

A report on the ReX exchange program Tilburg University, The Netherlands and the Natural Language and Information Processing (NLIP) Group at the Computer Laboratory, University of Cambridge, UK

By Sabine Buchholz (buchholz@kub.nl)

As not all readers of ;login: might be familiar with the research field of Computational Linguistics which forms the scientific background of this report about my four-months exchange stay at Cambridge University, UK, I will start by introducing some of the most important concepts. Computational Linguistics studies the combination of computers and natural, i.e. human, languages. It aims at developing and implementing models of how natural languages can be processed. Applications include text-to-speech, machine translation, question answering and natural language interfaces. A common subtask in many applications is parsing: determining the syntactic structure of a sentence. Although to a certain extent parsing can be done on the basis of knowledge about the part-of-speech (like verb, noun, preposition) of words, it is widely acknowledged that information about specific words (lexical knowledge), is advantageous. One of the most important pieces of lexical information is subcat(egorization), especially of verbs. This tells us for example that a verb like "give" preferably takes two complements (the ditransitive frame): "to give somebody something", whereas "invent" takes one complement (transitive): "to invent something" and "sleep" takes none (intransitive): "to sleep" but not "to sleep something". This information helps the parser in disambiguating sentences that would otherwise be ambiguous, like "She gave/invented Tim water." As parsers should be applicable to all kinds of texts, from all domains (for example for applications on the internet), and extensive subcategorization information is not readily available for all verbs, it can best be acquired automatically. It is this subcat acquisition problem that I worked on during my exchange to Cambridge.

Ted Briscoe is a reader in Computational Linguistics in the Natural Language and Information Processing Group (NLIP) group at Cambridge University. He had previously developed an automatic subcat acquisition system that works by parsing large amounts of texts (parsing based on part-of-speech information only), recording the frequency with which each frame occurs with each verb and filtering out combinations that did not occur sufficiently frequently (and are thus probably due to parser errors). Those verb-frame combinations that pass the filter, together with their associated frequencies (converted to probabilities) can subsequently be used for better probabilistic parsing.

To improve the performance of this last filtering step, PhD student Anna Korhonen developed a method for smoothing the acquired frequency distributions of new verbs by backing-off to semantically related known verbs, and for filtering based on the maximum likelihood estimation (MLE) of the resulting frequencies. As her method presupposes knowledge about semantic classes of verbs which is not easily available for all verbs, my task for the project was to explore alternative filtering approaches using machine learning.

I was a fourth-year PhD student in Tilburg, The Netherlands. As part of my research is on finding grammatical relations between verbs and their complements, which is related to parsing and subcat acquisition, I already knew several of Ted's and Anna's publications on the subject when Ted asked my supervisor Walter Daelemans whether one of Walter's students would be interested in the project. Walter had developed a machine learning algorithm called Memory-Based Learning (based on the k-Nearest Neighbor algorithm) which I also use for my thesis research, so I had the necessary background for the project and also liked the idea to spend some time in another foreign country. On the one hand, Cambridge with its beautiful and famous university, dating back to the 13th century, is very different from the "modern industrial city" Tilburg with its 75-year old university. On the other hand, everybody cycles there too, the landscape is conveniently flat and the city small, so I immediately felt at home. I arrived in August, which is a good time for getting to know the city, the river and the surroundings but a bad time for arriving at a university as half the staff is on holidays, conferences or summer schools. My first weeks were complicated by the fact that the entire computer laboratory, of which the NLIP group is a part, was moving to a new building at the western edge of the city (an event which should have happened long before I arrived but which had been postponed several times). So I started by (re) reading the available literature, most notably Anna's nearly finished thesis, and by discussing a lot with Ted and Anna. Once I got my own office and computer in the new building, I started to locate all of the corpus resources, acquisition system modules and evaluation software I had been reading about, and to use them myself. I also had a look at the source code, reviving my knowledge of Lisp, C and shell scripting on the way. I then worked on three topics:

After the subcat acquisition system has parsed a text, tokens of frames together with verbs can be extracted. These must then be classified into types of frames. For example "He invented the telephone" and "The telephone was invented" instantiated the same kind of (transitive) frame. I adapted the classifier to accomplish this passive-to-active conversion, so that all the tokens would contribute to the frequency count of their mutual type.

To evaluate how well the subcat acquisition performs, a so-called gold standard had been created manually. This means that some verbs were chosen at random, people looked at a representative number (mostly 300) of sentences in which these verbs occur, and noted how often they occur with which frames. Performance of the system is then computed in terms of precision (how many of the verb-frame pairs that the system proposes are also in the gold standard), recall (how many of the pairs in the gold standard are found by the system) and rank correlation (how similar is the order of pairs if gold standard and system results are ordered by frequency).
However, there are more types of frames that should be distinguished on theoretical grounds than the subcat acquisition system is able to do on the basis of part-of-speech information alone. Therefore the output of the system frequently was not a list of frames for each verb but a list of frame disjunctions.
These disjunctions complicate the computation of precision, recall and rank correlation. In addition, they made the results of evaluation of the machine learning experiments hard

to judge, as the learner tended to predict all possible disjunctions containing common frames. I therefore developed a variant of the classifier that is forced to return a list of single frames (no disjunctions). It is an open question what would be the best way to make such a forced decision. At the moment, the most general or frequent frame of a disjunction is chosen.

I used a supervised machine learning algorithm. This means that one part of the gold standard material was used for training the algorithm and another part for testing it. For each verb-frame pair acquired by the subcat acquisition system, the learner has to make a binary decision: keep it or reject it. As a first step, I had to create a machine learning instance for each such pair. Features of the instances correspond to pieces of information from system output or from external information sources (like the semantic classes used by Anna). I could then study the influence of various (combinations of) features on filter performance. After machine learning filtering, instances need to be converted back to the initial format of lexical entries. Results are that the influence of features depends heavily on the sort of verbs tested. In general, a combination of type of frame and observed frequency performs well, and adding additional information about semantic classes helps a little. For a special group of verbs however, the type of frame feature alone is sufficient and adding frequency information degrades performance. An experiment that still needs to be performed is to combine Anna's back-off smoothing with the machine learning filtering.

As I devoted much time into documenting my software during the last days of my exchange, research into the method can be continued after the official end of the exchange project. The documentation will form part of a larger technical report that should describe the subcat acquisition system and related modules. I will also use my new knowledge to make comparisons between the subcat acquisition system and parts of my thesis work.

I had a very pleasant and informative stay and wish to thank all the people who made this possible.

For more information about this exchange, please contact:

Dr. Sabine Buchholz
Computational Linguistics
Faculty of Arts
Tilburg University
PO Box 90153
5000 LE Tilburg
The Netherlands
S.Buchholz@kub.nl

Dr. Ted Briscoe
University of Cambridge Computer Laboratory
Pembroke St.

Cambridge CB2 3QG
United Kingdom
Ted.Briscoe@cl.cam.ac.uk

USENIX and Stichting NLnet jointly support the ReX program. For more information about ReX, see: http://www.usenix.org/about/rex.html.