CritiqueSamples

This page containts three critiques.

First Critique

Title: A Prototype Reading Coach that Listens
Authors: Jack Mostow, Steven F. Roth, Alexander G. Hauptmann, and Matthew Kane
Critique By: Veselin Stoyanov

Probably one of the reasons the paper was awarded the best paper award is the social implication and the practical applicability that the technology described in the paper has. It is not so often that you come across papers in computer science that have such a broad social implication. As the authors argue, even a partial solution to the problem that the paper attempts to solve would have a broad economic and social impact. As we become more and more dependent on Information Technology in the everyday life, the negative effect of illiteracy on members of the society is bound to increase, further increasing the need for solving the problem.

Another strength of the paper is the careful design of experiments and processing of the empirical data. Especially useful was the separation of the two hypotheses that the interventions are pedagogically effective and that they can be automated. Without separating the two experiments the data that the paper had gathered might have been inconclusive. The authors also designed the experiments carefully to offset possible random effects and carefully avoided the use of training data in the evaluation. In addition, the authors performed tests for statistical significance of the data, which help increase the confidence in the hypotheses that are presented.

From the introduction of the paper appears that the system is designed to combat illiteracy by teaching children to read. The authors further claim that the interventions are pedagogically effective. The hypothesis in section 2.1 (Pedagogical evaluation), however, is that the intervention would enable struggling readers to read and comprehend material significantly better compared to reading alone. It seems intuitive that the enhanced reading comprehension by using Emily will lead in the long run to learning to comprehend better alone. The authors do not present evidence that using Emily will help the children read better alone, however. Such an evaluation of the long term effect of using a technique as the one described in the paper may be useful as a part of future work, especially considering the danger of children becoming dependant on the system they are using for reading.

Although the paper presents improvements in the recognition part of the task compared to the predecessor of Emily, Evelyn, a possible drawback to the system is the low absolute level of speech recognition. Possible area for future work is to improve the level of correct recognition of errors by the recognition component (while preserving the current low level of false alarms). That part of the system, the lexical and language models of the speech recognizer lacks a theoretical background. The authors of the paper modified those two models of the recognizer in a way that produces the best empirical results, but it seems that the modifications do not follow a theoretical model. The authors themselves expected better results when using empirically observed in previous systems probabilities for the HMM of the recognizer. Contrary to the expectations, the empirical probabilities didn't work as well and the authors used "common sense" probabilities. It is worth experimenting in future work with using the probabilities observed empirically and using models from probabilistic theory to smooth out the probabilities in such a way as to offset the garden effect while possibly obtaining better probabilistic lexical and language models, which can lead to a better recognition level.

Another interesting area for future work is in the intervention part experimenting with the set of interventions and studying the effect of different intervention on comprehension. If the system is to be implemented in a real-world automated reading teacher, the effectiveness of interventions should be maximized. This aspect calls for carefully studying the interventions that different human teachers use in their pedagogical endeavors as well as evaluating research from Cognitive Studies and Psychology for cognitive models of reading comprehension and learning to read. After performing such a study a set of interventions can be designed and individual as well as groups of interventions in the set evaluated empirically as it was done in the paper and the set of most effective interventions (or rather the most effective set of interventions) can be chosen as the one implemented in the system. Here, the importance of testing the long term effects of learning to comprehend in addition to text comprehension while using the system is essential. Evaluating the interventions before designing the rest of the system is important because different interventions may be supported by different system designs. For instance, the system described in the paper does not allow for interruption of the reader in the middle of a sentence. In addition, the authors already argue in the paper that the analysis of the data can help make the interventions more effective as illustrated by the seldom use of rereading due to the help button. Of course, in designing the interventions the limitations of the recognizing component have to be considered as they appear to be the limiting component of the current system.

Finally, the result in the paper of the "potential" comprehension level being slightly lower than the "assisted" reading level is surprising especially compared to the cited result by (Curtis, 1980). The authors attribute that to the effect of lost if attention in the subjects due to the lack of a natural visual focus presented by a talking face. An interesting further investigation would be to perform the same experiment by either using a human to read the stories (ideally the person who's voice was used for the design of the system) or using graphics (such as simulated face) to perform the same experiment and evaluate the potential reading level. Such a study will have serious implications on areas such as Psychology and Human Computer Interaction. Furthermore, if the study shows that the speculation by the authors is correct that would suggest that the system can be possibly augmented with graphical aids that can help users to keep their attention on the task at hand.

Second Critique

Title: Integrating Multiple Knowledge Source to Disambiguate Word Sense: An Exemplar-Based Approach
Authors: Hwee Tou Ng and Hian Beng Lee
Critique By: Veselin Stoyanov

One of the strengths of the paper is the careful design of experiment and calculation of the results (at least for the first experiment). Authors average the results over multiple runs of randomly drawn test set for the first experiment and compute standard deviation in addition to the average. The results are compared historically with previously suggested algorithms.

One of the contributions of the authors is the creation of a large data set that can be used for evaluation of WSD algorithms. While being a useful tool, a few aspects of the new collection are questionable. First, the collection is annotated on the most frequently occurring and most ambiguous words in English. It does not become clear from the paper what qualifies as the most ambiguous words, but it appears that the most ambiguous words are the ones that have the most different meanings according to WordNet. It is not clear that the words with the most WordNet senses are the most ambiguous, since some senses may not even occur in the corpus.

Second, it appears somewhat unfair to evaluate algorithms only on the most frequently occurring words, since that means that more training data for the classifiers for those words is available. That puts at disadvantage algorithms that can learn classifications better from small number of examples.

Third, the estimate of 10-20% error seems somewhat arbitrary. Authors give no description of how this estimate was reached. It appears that agreement of 57% of the data is not an impressive result. However, I am not familiar with typical annotator agreement results for sense assignment, so I cannot judge the result. Nevertheless, a small subset of the data could have been overlapped between annotators to study agreement between different annotators, as high agreement between annotators should increase our confidence in the quality of the annotations used in the evaluation.

Lastly, authors talk about evaluating the algorithm on two separate test sets, a Brown corpus and a WSJ sets. It is not clear from the paper, but it appears that the algorithm has been trained on the entire corpus. It would be interesting to also evaluate a version trained only in the BC portion on the BC corpus and similarly for WSJ. It is not clear how the two test sets were selected and why multiple runs were not used and the results averaged as in the first experiment.

The algorithm that the paper proposes appears to be novel in its application to WSD and shows impressive results. A few steps in the design of the algorithm can be further elaborated on in future work. LEXAS uses a POS tagger and the morphological analyzer of WordNet. It is not clear how the performance of these two components affect the performance of the algorithm. While the authors make a good case for the POS tagger that typically achieves accuracy of about 96%, no such case is made for the morphological analyzer. Even in the case of good performance of the analyzer, an evaluation of the effect of the performance of the two external components on the performance of LEXAS might be useful as it can guide where efforts for further improving the algorithm should be concentrated.

In selecting the features constituting the frequently co-occurring words, the local collocation, and the verb-object syntactic relation, authors select based on conditional probability. From the description of the algorithm, it appears that a more desirable measure could be mutual information, since sometimes words with slightly lower conditional probability may be a good predictor of the sense of the word under consideration. For instance, a word can be a good indicator for senses 1 and 2 and indicate near 0 likelihood for senses 3 to n. Conditional probability will ignore such a word as a good feature, while in reality including such a feature might be useful.

Additionally, the choice of the nine features for local collocations appears to be arbitrary. Authors do not justify the particular choice of the nine features and not let's say -3 : 3 for instance. A more reasonable approach would be to consider all combinations of features up to a given length and estimate from the training set which features are the most useful ones.

In future work authors may consider implementing a higher-precision method for determining the verb-object syntactic relation. An approach using a syntactic parser may enable a higher precision and allow different syntactic relations to be considered. Additionally, syntactic parsing technique could allow determining the verb-object relation when the noun under consideration is not the first argument of the verb.

Finally, the algorithm used for classification is k nearest neighbors with k = 1. It would be interesting to evaluate the same algorithm with different values for k and compare the results. Furthermore, it would be interesting to compare the performance of KNN with other algorithms known to often outperform KNN such as Naive Baise and SVM. The problem at hand can be thought of as multi-label classification.

Third Critique

Title: How Reliable are the Results of Larg-Scale Information Retrieval Experiments?
Author: Justin Zobel
Critique By: Oren Kurland

The goal of the paper is to test how reliable are the performance measurement techniques used in the TREC (and generalize to large-scale IR systems) and how reliable the pooling method is as a mean to determine which documents are the relevant ones.

Estimating the validity of the precision measurements

1. The results of precision measurements in the TREC are proved to be valid and therefore the TREC can be truly regarded as a reliable tool to differentiate between the precision performance of different IR systems. Moreover, the techniques to measure precision could be safely (without fear of unfairness or invalidity) deployed by those who measure the performance of large-scale IR systems.
2. The shown result that as long as the difference between the systems' (retrieval techniques) performance is statistically significant the type of the measurement used is relatively unimportant, is important for the question of which measurement to use when testing an IR system. This result also indicates the high importance of using significance testing (which is easy to implement) when testing the improvements one makes to a specific IR system. Nevertheless, the shown result is only partially proved since the correlation between the different measures wasn't tested by the author.
3. Although the author shows that Wilcoxon's test has the best discriminating results for 25 samples (in comparison to t-test and ANOVA), he/she doesn't mention that Wilcoxon's n(s/r) parameter is probably greater than 10 and therefore Wilcoxon's test in this case approximates normal distribution, the same as the t-test does (with sample size of ~30). So the conclusion of choosing Wilcoxon's test is not obvious and the author should have compared the statistical tests with respect to a varying number of queries (10,20 for un normal distribution and more than 30 for the normal one).

The Pooling method

1. An important result is that the use of a measurement depth larger than the pool depth is unjustified because it introduces a great deal of uncertainty to the results and the TREC organizers should take it into account.
2. The pool's depth of 100 which is used in the TRECK is adequate for measurement of precision but not for recall since only 50%-70% of the relevant documents in the corpus are found using the pooling technique.
3. The suggested (incremental) algorithm for pooling is important in the sense that it will enable those who evaluate the performance of large-scale IR systems to both economize the manual efforts of determining the relevant documents (because the algorithm may converge to relatively small depths) and to obtain more reliable results when determining the set of relevant documents in the entire collection.