Evaluating Evaluation Metrics for Ontology-Based Applications: Infinite Reflection

Diana Maynard, Wim Peters, Yaoyong Li

Dept of Computer Science, University of Sheffield, Sheffield, UK {diana,wim,y.li}@dcs.shef.ac.uk

Abstract content

In this paper, we discuss methods of measuring the performance of ontology-based information extraction systems. We focus particularly on the Balanced Distance Metric (BDM), a new metric we have proposed which aims to take into account the more flexible nature of ontologically-based applications. We first examine why traditional Precision and Recall metrics, as used for flat information extraction tasks, are inadequate when dealing with ontologies. We then describe the Balanced Distance Metric (BDM) which takes ontological similarity into account. Finally, we discuss a range of experiments designed to test the accuracy and usefulness of the BDM when compared with traditional metrics and with a standard distance-based metric. The main problem is one of infinite reflection: namely that we need methods to evaluate an evaluation metric, but how do we know that these methods themselves are pertinent? In this paper we describe and discuss a variety of methods to evaluate the metrics, covering comparison of experimental results with human intuition, range of evaluation covered, comparison of performance on different kinds of systems, and scalability when applied to different granularities of ontology. We show, amongst other things, that the BDM is in many ways an improvement over the traditional Precision and Recall metrics for evaluating ontology-based information extraction tasks.

Traditionally, information extraction systems have been evaluated using Precision and Recall, which classifies each entity returned by the system as either correct or incorrect. However, this is not sufficient for ontology-based information extraction, because the distinction between correct and incorrect is more fuzzy: if an answer is closely related to the correct answer, then some credit should be given for an "almost correct" answer, rather than simply classifying it as wrong. For example, with traditional information extraction, either something is a Person or it is not, but with ontology-based information extraction, something might be classified as a Researcher rather than an Academic, which could often be seen as partially correct given that both are subclasses of Person. So a metric which classifies the correctness of an answer based on its semantic proximity to the real answer should give us a fairer indication of the performance of the system. Other existing cost-based or distance-based metrics, such as Learning Accuracy (LA) (Hahn and Schnattinger, 1998), have some flaws such as not taking into account the density of the hierarchy, and in the case of LA, being asymmetrical.
The BDM computes semantic similarity between two semantic annotations of the same token in a document. The metric has been designed to replace the traditional "exact match or fail" metrics with a method which yields a graded correctness score by taking into account the semantic distance in the ontological hierarchy between the compared nodes (Key and Response). The final version of the BDM is a slightly improved version of the original (Maynard, 2005), which did not take the branching factor into account (as described below).
The BDM is computed on the basis of the following measurements:

- CP = the shortest length from root to the most specific common parent, i.e. the most specific ontological node subsuming both Key and Response)
- DPK = shortest length from the most specific common parent to the Key concept

- DPR = shortest length from the most specific common parent to the Response concept
- n1: average chain length of all ontological chains containing Key and Response.
- n2: average chain length of all ontological chains containing Key.
- n3: average chain length of all ontological chains containing Response.
- BR: the branching factor of each relevant concept, divided by the average branching factor of all the nodes from the ontology, excluding leaf nodes.

The complete BDM formula is as follows:

$$BDM = \frac{BR(CP/n1)}{BR(CP/n1) + (DPK/n2) + (DPR/n3)} \quad (1)$$

The BDM itself is not sufficient to evaluate our populated ontology, however, because we need to preserve the useful properties of the standard Precision and Recall scoring metric. Our APR metric (Augmented Precision and Recall) combines the traditional Precision and Recall with a cost-based component (namely the BDM). We thus combine the BDM scores for each instance in the corpus, to produce Augmented Precision, Recall and F-measure scores for the annotated corpus, calculated as follows:

$$AP = \frac{BDM}{n + Spurious} \; and \; AR = \frac{BDM}{n + Missing} \quad (2)$$

while F-measure is calculated from Augmented Precision and Recall as:

$$F - measure = \frac{AP * AR}{0.5 * (AP + AR)} \quad (3)$$

When evaluating the BDM as a new metric, we suggest the following criteria as proposed by (King, 2003) for evaluation metrics in general. A metric should:

- reach its highest value for perfect quality;
- reach its lowest value for worst possible quality;
- be monotonic;
- be clear and intuitive;
- correlate well with human judgement;
- be reliable and exhibit as little variance as possible;
- be cheap to set up and apply;
- be automatic.

In this paper we show how the BDM fulfils these criteria. Below we summarise some of the experiments we have carried out to investigate the validity of the BDM as a new metric.

In order to test the effectiveness of the BDM, we carried out some experiments to compare it with 2 other metrics, Learning Accuracy and the flat traditional measure, using the Hieron (Li et al., 2006) and SVM (Cristianini and Shawe-Taylor, 2000; Li et al., 2005) learning algorithms for OBIE. The SVM algorithm was a flat classification in which the structure of concepts in the ontology was ignored, while the Hieron algorithm was based on hierarchical classification that exploits the structure of concepts.

Both the BDM_F_1 and LA_F_1 are higher than the flat_F_1 for the two algorithms, reflecting the fact that the latter only counts the correct classifications, while the former two not only count the correct classifications but also the incorrect ones. However, the difference for the Hieron is more significant than that for the SVM, demonstrating an important difference between the two methods — the SVM based method just tries to learn a classifier for one concept as well as possible, while the Hieron based method not only learns a good classifier for each individual concept but also takes into account the relations between the concepts in the ontology during the learning.

In terms of the conventional flat_F_1, the Hieron was slightly better than the SVM. However, if the results are measured by using the ontology-sensitive measure BDM_F_1 or LA_F_1, we can see that the Hieron performed significantly better than the SVM. Clearly, the ontology-sensitive measures such as the BDM_F_1 and LA_F_1 are more suitable than the conventional flat_F_1 to measure the performance of an ontology-dependent learning algorithm such as

Hieron. We also looked at human judgement to see how well the BDM performed compared with our expectations of similarity. Further details of human judgement experiments will be given in the full paper, but we discovered that on a small-scale experiment, the BDM did appear to correlate very well with human judgement on similarity. For example, the LA gives a score of 1 (i.e. 100% correctness) for the entity "Sochi" being annotated by the system as "Location" when the correct answer should be the more precise concept "City". The BDM gives a score of 0.72 (i.e. 72% correctness) which is much more appropriate to something that we can consider as a close match.

Another experiment performed was to compare these OBIE learning algorithms in GATE with the KIM system (Popov et al., 2004), using Precision and Recall versus BDM. Interestingly we found that the difference in performance between the two systems is much smaller with the BDM metric, reflecting very well the fact that KIM finds many entities but does not always classify them absolutely correctly. Such minor misclassifications are heavily penalised with traditional metrics but much less heavily penalised with the BDM. This is a more accurate reflection of the system's performance because in many cases, such minor misclassifications are not so important.

It is also important to measure how scalable a new evaluation metric is. Specifically, we investigate how the BDM measures up to other metrics when the ontology is collapsed or expanded in various ways, and what happens with smaller or larger ontologies. We therefore performed some experiments to measure this, by comparing annotation systems using different metrics on 3 different versions of the Proton ontology, which we created specifically for the experiment. PTop was based on the concept levels of the ontology, and was created by just keeping the concepts with the "ptop" tag in the original Proton ontology, i.e. the uppermost concepts. Other concepts in Proton were mapped to the nearest ancestor concept, i.e. "ptop". This reduced the number of concepts from 272 to 25. Link-1 was based on the link characteristics. For each node in the ontology, if it was the only child concept of its parent, then the node was collapsed with its nearest ancestor concept with more than one child node. This reduced the ontology size from 272 to 244 concepts. We then compared 4 different metrics on the annotations: flat (traditional Precision and Recall), distance (a measure based on very simple hierarchical distance), Learning Accuracy, and the BDM. More details will be given in the full paper, but the main conclusions to be drawn here were that all three hierarchical measures are better than conventional measures for evaluating ontology-based annotation, and that the BDM is less sensitive to ontology size, and the only metric to reflect ontology density.

In the final paper, we will give a more in-depth analysis of the experiments performed to test the validity of the BDM and to compare it with other potential metrics. For example, we will show the results of human correlation experiments, as mentioned above. In summary, the paper presents a metric for evaluating ontology-based information extraction and a set of experiments designed to test how suitable this metric really is for the task in question. This can be generalised to a more widespread method of evaluating evaluation metrics.

1. References

N. Cristianini and J. Shawe-Taylor. 2000. *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

U. Hahn and K. Schnattinger. 1998. Towards text knowledge engineering. In *Proc. of 15th National Conference on Artificial Intelligence (AAAI-98)*, pages 524–531, Menlo Park, CA. MIT Press.

M. King. 2003. Living up to standards. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing*, Budapest, Hungary.

Y. Li, K. Bontcheva, and H. Cunningham. 2005. SVM Based Learning System For Information Extraction. In M. Niranjan J. Winkler and N. Lawerence, editors, *Deterministic and Statistical Methods in Machine Learning*, LNAI 3635, pages 319–339. Springer Verlag.

Y. Li, K. Bontcheva, and H. Cunningham. 2006. Perceptron-like learning for ontology based information extraction. Technical report, University of Sheffield, Sheffield, UK.

D. Maynard. 2005. Benchmarking ontology-based annotation tools for the semantic web. In *UK e-Science Programme All Hands Meeting (AHM2005) Workshop "Text Mining, e-Research and Grid-enabled Language Technology"*, Nottingham, UK.

B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. 2004. KIM – Semantic Annotation Platform. *Natural Language Engineering*.