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

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.

1. Introduction

Traditionally, applications such as information extraction (IE) are evaluated using Precision, Recall and F-Measure. These metrics give us a binary decision of correctness for each entity in the text, i.e. by comparing the key (gold standard) and system responses, they classify the result as either right or wrong in each case. Ontology-based information extraction systems attempt to classify entities in a more scalar fashion, as there are many different categories to which an entity can be assigned, and the distinction between these categories is much less clearcut. Traditional entity classes do not subsume each other, whereas in an ontology, there are subclass and superclass categories to consider, so the distinction between right and wrong is more blurred. In traditional IE, an element identified as a Person is either correct or incorrect (measured by Precision), and elements which should be identified as Person are either identified or not (measured by Recall). When making an ontological classification, however, the distinction is more fuzzy. For example if we misclassify an instance of a Researcher as a Lecturer, we are clearly less wrong than missing the identification (and classification) altogether, and we are also somehow less wrong than if we had misclassified the instance as a Location. Credit should therefore be given for partial correctness. Traditionally, this is sometimes achieved by allocating a half weight to something deemed partially correct, but this is still insufficient to give a proper distinction between degrees of correctness. We therefore adopt an approach based on similarity between Key (the gold standard) and Response (the output of the system), known as BDM (Maynard, 2005; Maynard et al., 2006).

In this paper, we aim to evaluate how useful the BDM is as a measure of the performance of an ontology-based Information Extraction (OBIE) system. Of course, how to evaluate an evaluation metric is not obvious, but there are some general guidelines proposed by (King, 2003) for evaluation metrics that we try to follow. 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.

We aim in this paper to show how the BDM fulfils these criteria, describing some experiments we have carried out to investigate its validity.

2. A Distance-Based Metric for Evaluation

As discussed in Section 1., 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. By this we mean that comparing two concepts in the ontology gives different results depending on which one is the Key and which is the Result. Given that we trying to compute the similarity between two concepts, it seems rather odd and unintuitive that this should be the case.

The BDM computes semantic similarity between two 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 an improved version of the original BDM described in (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, 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:

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

3. Evaluation of Ontology-Based Information Extraction Metrics

We performed various experiments to test the validity of the BDM metric. First, we compared it with two other metrics: a flat metric and another similarity-based metric, the LA. We also looked at how it performed on two different learning algorithms for IE: one which is based on a flat classification (SVN) and one which is based on a hierarchical classification (Hieron). Our second experiment compared two different IE systems (GATE and KIM) using Precision and Recall, and the APR, in order to see how they two systems compared using the two different metrics. The idea behind this was to see if using BDM revealed any differences between the two systems that traditional metrics did not. The third experiment looked specifically at the scalability of the metric and how it performed on different densities and sizes of ontology.

For the evaluations, we used the semantically annotated OntoNews corpus (Peters et al., 2005) as a gold standard. This consists of 292 news articles from three news agencies (The Guardian, The Independent and The Financial Times), and covers the period of August to October, 2001. The articles belong to three general topics or domains of news gathering: International politics, UK politics and Business. The ontology used in the generation of the ontological annotation process was the PROTON ontology[1], which has been created and used in the scope of the KIM platform[2] for semantic annotation, indexing, and retrieval (Kiryakov et al., 2004). The ontology consists of around 250 classes and 100 properties (such as partOf, locatedIn, hasMember and so on). PROTON has a number of important properties: it is domain-independent, and therefore suitable for the news domain, and it is modular (comprising both a top ontology and a more specific ontology).

3.1. Experiments with OBIE

The aim of the first set of experiments was, on the one hand, to evaluate a new learning algorithm for OBIE, and, on the other hand, to compare the different evaluation metrics (LA, flat traditional measure, and the BDM).

The OBIE algorithm learns a Perceptron classifier for each concept in the ontology. Perceptron (Rosenblatt, 1958) is a simple yet effective machine learning algorithm, which forms the basis of most on-line learning algorithms. Meanwhile, the algorithm tries to keep the difference between two classifiers proportional to the cost of their corresponding concepts in the ontology. In other words, the learning algorithm tries to classify an instance as correctly as it can. If it cannot classify the instance correctly, it then tries to classify it with another concept with the least cost associated with it relative to the correct concept. The algorithm is based on the Hieron, a large margin algorithm for hierarchical classification proposed in (Dekel et al., 2004). See (Li et al., 2006) for details about the learning algorithm and experiments.

We experimentally compared the Hieron algorithm with the SVM learning algorithm (see e.g. (Cristianini and Shawe-Taylor, 2000)) for OBIE. The SVM is a state of the art algorithm for classification. (Li et al., 2005) applied SVM with uneven margins, a variant of SVM, to the traditional information extraction problem and achieved state of the art results on several benchmarking corpora. In the application of SVM to OBIE, we learned one SVM classifier for each concept in the ontology separately and did not take into account the structure of the ontology. In other words, the SVM-based IE learning algorithm was a flat classification in which the structure of concepts in the ontology was ignored. In contrast, the Hieron algorithm for IE is based on hierarchical classification that exploits the structure of concepts.

As the OntoNews corpus consists of three parts (International politics, UK politics and Business), for each learning algorithm two parts were used as training data and another part as test data. Note that although the tripartition of the corpus indicates three distinct and topically homogeneous parts of the corpus, these parts are used as training and testing data for the comparison of different algorithms, and not

[1] http://proton.semanticweb.org
[2] http://www.ontotext.com/kim

their performance. For this purpose, semantic homogeneity does not play a role.

For each experiment we computed three F_1 values to measure the overall performance of the learning algorithm. One was the conventional micro-averaged F_1 in which a binary reward was assigned to each prediction of instance — the reward was 1 if the prediction was correct, and 0 otherwise. We call this flat_F_1 since it does not consider the structure of concepts in the ontology. The other two measures were based on the BDM and LA values, respectively, which both take into account the structure of the ontology.

	flat_F_1	BDM_F_1	LA_F_1
SVM	73.5	74.5	74.5
Hieron	74.7	79.2	80.0

Table 1: Comparison of Hieron and SVM for OBIE

Table 1 presents the experimental results for comparing the two learning algorithms SVM and Hieron. We used three measures: conventional micro-averaged flat_F_1 (%), and the two ontology-sensitive augmented F_1 (%) based respectively on the BDM and LA, BDM_F_1 and LA_F_1. In this experiment, the International-Politics part of the OntoNews corpus was used as the test set, and the other two parts as the training set.

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 Hieron is more significant than that for SVM, demonstrating an important difference between the two methods — the SVM based method just tried to learn a classifier for one concept as well as possible, while the Hieron based method not only learned a good classifier for each individual concept but also took into account the relations between the concepts in the ontology during the learning.

In terms of the conventional flat_F_1, the Hieron algorithm performed 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 Hieron performed significantly better than 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.

In order to analyse the difference between the three measures, Table 2 presents some examples of entities predicted incorrectly by the Hieron based learning system, their key labels, and the similarity between the key label and predicted label measured respectively by the BDM and the LA. Note that in all cases, the flat measure produces a score of 0, since it is not an exact match.

All the concepts and their relations involved in Table 2 are illustrated in Figure 1, which presents a part of the PROTON ontology. This ontology section starts with the root node *Thing*, and has 10 levels of concepts with *TVCompany* as the lowest level concept. Note that the graph does not show all the child concepts for most of the nodes presented.

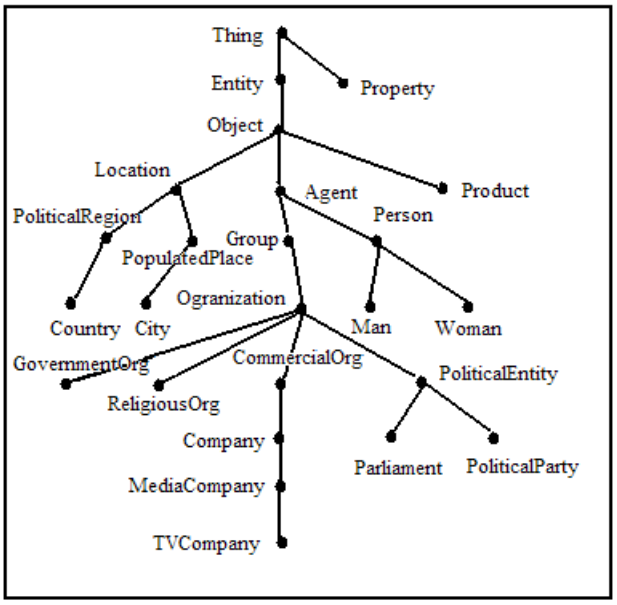


Figure 1: Subset of the PROTON ontology

The conventional flat measure assigned each case a zero similarity because the examples were misclassified and the measure does not consider the structure of labels. On the other hand, both the LA and BDM take into account the structure of labels and measure the degree of a misclassification based on its position in the ontology. Hence they assign a non-zero value to a misclassification in most cases. Note that zero would be assigned in the case where the MSCA is the root node. In our experiments, all the concepts used were below the node "Entity" and so we used its immediate upper node "Thing" as root[3]. This meant that CP (the depth of the MSCA) was always at least 1, and hence there is no zero value for BDM or LA in our experiments. This is because we consider that if an entity's instance is recognised but with the wrong type, the system should have a non-zero reward because it at least recognised the instance in the first place. However, this could be changed according to the user's preference.

However, BDM and LA adopt different mechanisms in consideration of the ontology structure. In particular, the LA assigns the maximal value 1 if the predicted label is an ancestor concept of the key label, regardless of how far apart the two labels are within the ontological chain. In contrast, the BDM takes into account the similarity of two concepts in the ontology and assigns a distance-dependent value. The difference is demonstrated by the examples in the table. For example, in the Proton ontology, the predicted label *Organization* is the parent concept of the key label *GovernmentOrganization* in the second example, and in the third example the same predicted label *Organization* is 4 concepts away from the key label *TVCompany*. Hence, the BDM value of the second example is higher than the

[3]"Thing" subsumes both "Entity" and "Property"

No.	Entity	Predicted label	Key label	BDM	LA
1	Sochi	Location	City	0.724	1.000
2	Federal Bureau of Investigation	Organization	GovernmentOrganization	0.959	1.000
3	al-Jazeera	Organization	TVCompany	0.783	1.000
4	Islamic Jihad	Company	ReligiousOrganization	0.816	0.556
5	Brazil	Object	Country	0.587	1.000
6	Senate	Company	PoliticalEntity	0.826	0.556
7	Kelly Ripa	Man	Person	0.690	0.667

Table 2: Examples of entities misclassified by the Hieron based system

BDM value of the third example. In the first example, the predicted label *Location* is 3 concepts away from the key label *City* but its BDM value is lower than the corresponding value in the third example, mainly because the concept *Location* occupies a higher position in the Proton ontology than the concept *Organization*. Similarity is thus lower because higher concepts are semantically more general, and therefore less informative.

Another difference between the BDM and LA is that the BDM considers the concept densities around the key concept and the response concept, but the LA does not. The difference can be shown by comparing the fourth and the sixth examples. They have the same predicted label *Company*, and their key labels *ReligiousOrganization* and *PoliticalEntity* are two sub-concepts of *Organization*. Therefore, the positions of the predicted and key labels in the two examples are very similar and hence their LA values are the same. However, their BDM values are different — the BDM value of the fourth example is a bit lower than the BDM value of the sixth example. This is because the concept *PoliticalEntity* in the sixth example has two child nodes but the concept *ReligiousOrganization* in the fourth example has no child node, resulting in different averaged lengths of chains coming through the two concepts.

The BDM value in the fifth example is the lowest among the examples, mainly because the concept *Object* is in the highest position in the ontology among the examples. These differences in BDM scores show the effects of the adoption of chain density and branching factor as penalty weights in the computation of the score. These reflect the level of difficulty associated with the selection of a particular ontlogical class relative to the size of the set of candidates.

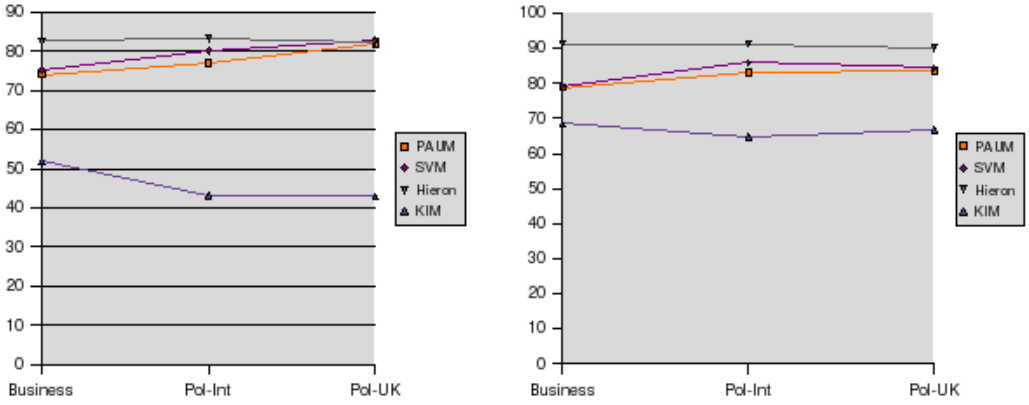
3.2. Comparison of GATE and KIM using BDM

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. We used the same set of texts and ontology as for the previous experiments; however, the articles in OntoNews were divided into three subsets according to the article's theme, namely business, international politics and UK politics. These sets contained 91, 99 and 100 articles, respectively. The results for traditional F-measure and AF measure (using BDM) on each of these sets and using each system are shown in Figure 2. We can see that with both metrics, GATE outperforms KIM. Interestingly, we found that the difference in performance between the two systems is much smaller with the BDM metric than with the traditional metric. This reflects very well the fact that due to the algorithms used in the IE process, KIM finds many correct instances of entities but does not always classify them absolutely correctly. GATE, on the other hand, tends to either find and classify the instance completely correctly, or not find it at all. 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.

3.3. Scalability of BDM

It is also important to measure how scalable a new evaluation metric is. Specifically, we investigated 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.

Space constraints mean we cannot give full details here, but the experiments enabled us to draw various conclusions. Unsurprisingly, all three hierarchical measures are better than conventional measures for evaluating ontology-based annotation. The BDM is less sensitive to ontology size than LA, because it considers normalisation with respect to ontology size, which is important as the ontology gets bigger. BDM is also the only one which reflects ontology density; the others only reflect size. We propose to do some further experiments to highlight this. With a very shallow ontology, it actually makes little difference which of these three measures is used. The larger (and deeper) the ontology, the more difference it makes which method we use. We would expect that as the ontology increases in size and complex-



(a) F-measures

(b) AF-measures

Figure 2: Comparison of GATE and KIM

ity, the more important the choice of metric is. Of course, shallowness of the ontology is not the only factor related to its size: properties, labels, or other ontology components also play a part. However, since they are not used for the ontology-based information extraction component, we are not interested here in evaluating these factors separately.

4. Discussion and Future work

The initial observation from our experiments is that binary decisions are not good enough for ontology evaluation, when hierarchies are involved. Both the BDM and LA metrics are more useful than a distance-based or flat metric when evaluating information extraction based on a hierarchical rather than a flat structure. From a human perspective, the BDM appears to perform better than the LA in that it reflects a better error analysis in certain situations. In terms of scalability, we see that the BDM appears robust when dealing with different sizes and densities of ontology, although we have only conducted a small-scale experiment so far. In contrast to other metrics, it reflects ontology density as well as size.

We also found an interesting feature of the BDM over a traditional flat metric, in that it enabled better distinction of some kinds of IE system. By penalising minor misclassifications less heavily than than completely undiscovered or unclassified entities, it reflects better the distinction between systems which find many entities but make minor mistakes in their classification, and systems which fail to actually find many entities or make major classification mistakes.

This paper has shown the usefulness of evaluation metrics like the BDM in ontology population. The BDM is solely based on the structure of the ontology to measure the similarity of concepts in the ontology. However, since the ontology can have semantic interpretation based on description logic, we could also measure concept similarity in the ontology using the underlying description logic. This kind of semantic similarity would make more sense than the structure-based measure, for a complex ontology containing different type-of relations. Future work will investigate this approach.

5. References

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

O. Dekel, J. Keshet, and Y. Singer. 2004. Large Margin Hierarchical Classification. In *Proceedings of the 21st International Conference on Machine Learning (ICML-2004)*, Canada.

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.

A. Kiryakov, B. Popov, D. Ognyanoff, D. Manov, A. Kirilov, and M. Goranov. 2004. Semantic annotation, indexing and retrieval. *Journal of Web Semantics, ISWC 2003 Special Issue*, 1(2):671–680.

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, W. Peters, and Y. Li. 2006. Metrics for evaluation of ontology-based information extraction. In *WWW 2006 Workshop on "Evaluation of Ontologies for the Web" (EON)*, Edinburgh, Scotland.

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.

W. Peters, N. Aswani, K. Bontcheva, and H. Cunningham. 2005. Quantitative Evaluation Tools and Corpora v1. Technical report, SEKT project deliverable D2.5.1.

B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. 2004. KIM – A semantic platform for information extraction and retrieval. *Natural Language Engineering*, 10:375–392.

F. Rosenblatt. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.