Proposal for Evaluating Ontology Refinement Methods

Enrique Alfonseca* **and Suresh Manandhar†**

* Departamento de Ingeniería Informática, Universidad Autónoma de Madrid
28049 Madrid, Spain
Enrique.Alfonseca@ii.uam.es

† Computer Science Department, University of York
York YO10 5DD, United Kingdom
suresh@cs.york.ac.uk

Abstract

Ontologies are a tool for Knowledge Representation that is now widely used, but the effort employed to build an ontology is still high. There are a few automatic and semi-automatic methods for extending ontologies with domain-specific information, but they use different

Approach	Method	Ontology	Corpus
Hearst (1992)	Det.	WordNet (Miller, 1995)	Grolier's Academic American Encyclopedia
Kietz et al. (2000)	Det.	GermaNet	corporate intranet
Alfonseca and Manandhar (2002)	Det.	WordNet	*The Lord of the Rings* (Tolkien, 1968)
Hastings (1994)	Non-det.	LINK hierarchy (Lytinen, 1991)	newswire articles
Hahn and Schnattinger (1998)	Non-det.	KL-ONE Terminological Knowledge Base (Woods and Schmolze, 1992)	I.T. magazines

Table 1: Comparison of different approaches for General Named Entity Identification

Hereon, we shall focus on the OR sub-step that consists in extending an ontology with new concepts, a task that Alfonseca and Manandhar (2002) called *General Named Entity Identification*. We can distinguish two subtasks:

- Locating the relevant new terms. For example, one can consider that the relevant terms for a domain are those that have a higher frequency in any text from that domain than in general-purpose texts.
- Placing them into the ontology, for instance, by indicating which are their maximally specific generalisations, amongst the concepts that are already inside the ontology.

We have classified reported work in this field in two main groups: deterministic and non-deterministic systems.

Deterministic systems are those that provide, for each unknown concept, one or several generalisations taken from the ontology, all of which are supposedly correct.

One of these systems, described by Hearst (1998), extended the WordNet lexical ontology (Miller, 1995). Using the standard terminology, when a concept a is a generalisation of a concept b, we say that a is a hypernym of b and that b is a hyponym of a. The approach followed by Hearst consists in finding regular-expression patterns from free texts by looking at pairs of (hypernym, hyponym) that co-occur in the same sentence, and then these patterns are used to learn new relations for extending WordNet. For example, the sentence (1) can be used to find that the pattern `such NPs as {NP,}* NP` usually states a hypernymy relation. However, he notes that these extracted relations contain a large number of mistakes.

(1) ...works by such authors as Herrick, Goldsmith and Shakespeare...

Kietz et al. (2000) applied similar hand-coded patterns for extending GermaNet (a German equivalent of WordNet) with concepts from a corporate intranet, and quantified the error rate in 32%. Therefore, there are two main drawbacks that have to be settled:

1. Unknown concepts that never appear in one of the expected patters cannot be classified.
2. The high error rate implies that it is necessary that a user validates the program output.

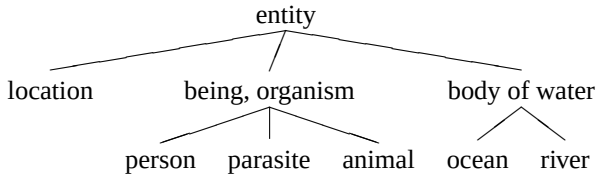
We recently described other deterministic algorithm to extend an ontology with domain-specific concepts extracted from specific texts (Alfonseca and Manandhar, 2002). Our system performs a top-down search through the ontology, selecting at each step the specialisation whose context words are more similar to the context words of the new concepts. This algorithm has been applied to extend WordNet with concepts extracted from *The Lord of the Rings* (Tolkien, 1968).

Non-deterministic systems, on the other hand, provide a set of likely candidate hypernyms amongst which there are some that are correct.

On of such systems, Camille, was built by Hastings (1994). In Camille, there are some concept ontologies for nouns and verbs about the terrorist domain, and the verbs are annotated with selectional preferences, e.g. the object of *arson* is known to be a *building*, and the object of *kill* is known to be an *animate being*.

If we have an unknown word u, initially, every concept in the ontology is a possible hypernym, i.e. the hypothesis space is the whole ontology. If, for instance, u was found being the direct object of *arson*, we would have evidence in favour of it being classified as a building, whilst at the same time *animated being* and all its specialisations can be ruled out from the hypothesis space. Finally, the set of resulting hypotheses is provided as result. A very similar approach was taken by Hahn and Schnattinger (1998). He used an ontology about electronic devices, and the constraints were as well about verbal selectional restrictions.

The difference between non-deterministic and deterministic systems is that the first provide the whole set of hypotheses that could be valid, from the evidence given in the text corpora, and do not try to guess which ones of these hypotheses are correct and which are incorrect. On the other hand, deterministic systems such as the one described by



u_j = lice
g_j = {parasite, animal}
c_i = lice
h_{i1} = {parasite, animal}
h_{i2} = {parasite, *person}
h_{i3} = {*location, *body of water}

Figure 1: Example of taxonomy, an unknown relevant concept u_j, its correct generalisations g_j and the generalisations proposed by three hypothetical algorithms h_{ik}.

Alfonseca and Manandhar (2002) sometimes have to do a *wild guess* when the evidence from the texts is scant.

Table 1 shows a summary of the related work, the ontologies used and the corpora from which they have extracted the new concepts.

3. Task description and settings

Let us suppose that we have a set of domain-specific documents D, containing some unknown concepts and instances $U = \{u_1, u_2, ..., u_n\}$, and an ontology O.
General Named Entity (GNE) Identification is the task that consists in finding, for every unknown concept or instance u_j found in a text, its maximally specific generalisations $\{g_1, g_2, ..., g_n\}$.

As can be observed, the task is similar to the IE task *Named Entity Identification*, in which unknown words have to be classified as people, locations, organisations, or any of a set of pre-defined classes. GNE identification is a more ambitious task, where the classes in which unknown words have to be classified are not specified beforehand; instead, these classes are organised as an ontology, and the classification system has to be able to handle different ontologies containing many possible kinds of information.

To properly compare ontology learning algorithms, we need to fix previously the training and test data, and a suitable evaluation metric.

3.1. Training data

The learning algorithms will most likely need two resources:

- **An existing ontology**. We have chosen WordNet 1.7, because there is no consensus in the existing literature, and WordNet is one of the most widely used.
- **A text collection** that can be used either by automatic procedures or to test hand-crafted methods to train the system. For example, Hearst (1992) used as training data the texts where he looked for co-occurring pairs of hypernyms and hyponyms, in order to find the word patterns. In the approach taken by Alfonseca and Manandhar (2002), the training data is used to generate, for every concept in the ontology, the set of context words that can appear in its neighbourhood. Those sets of context words will be compared to the context of new concepts in order to decide how to classify and introduce them into the ontology.

  Ideally, the text collection need to be fixed so different algorithms can be compared objectively, but given the vastness of the Internet it is plausible that fixing the document bank may not be that essential, if search engines are used to find relevant documents on Internet.

3.2. Test data

The ideal properties of the test data are the following:

- It must be domain-specific.
- It must contain concepts and instances not present in WordNet, so they can be learnt.

We have annotated two collections of texts to be used as test corpora: a portion of the *Wall Street Journal corpus* from the Penn Treebank (Marcus et al., 1993), about the economics domain, and Homer's *The Iliad*, a mythological text. Both are easily available for research purposes, and the first one has the added value that it has been used as benchmark corpus for many other tasks in Natural Language Processing.

3.3. Evaluation metrics

Let us suppose that we have a set of unknown concepts that appear in the test set and are relevant for an specific domain: $U = \{u_1, u_2, ..., u_n\}$. A human annotator has specified, for each unknown concept u_j, its maximally specific generalisations from the ontology: $G_j = \{g_{j,1}, ..., g_{j,m_j}\}$.

Let's suppose that an algorithm decided that the unknown concepts that are relevant are $C = \{c_1, c_2, ..., c_l\}$. For each c_i, the algorithm has to provide a list of maximally specific generalisations from the ontology: $H_i = \{h_{i,1}, h_{i,2}, ..., h_{i,p_i}\}$.

For illustration, let us consider the ontology in Figure 1. Let us suppose that the word *lice*, appearing in some domain-specific texts, is relevant enough, and therefore a human annotator has labelled it as u_i and has decided that its maximally specific generalisations are those in the set g_i. Let us suppose that three different automatic classifiers have also decided that it is a relevant concept, have annotated it as c_j and have chosen as generalisations the sets h_{i1}, h_{i2} and h_{i3}, respectively. We need evaluation metrics that show that the first algorithm is better than the second, which is itself better than the third one.

The following metrics have been taken, with small modifications, from Hastings (1994).

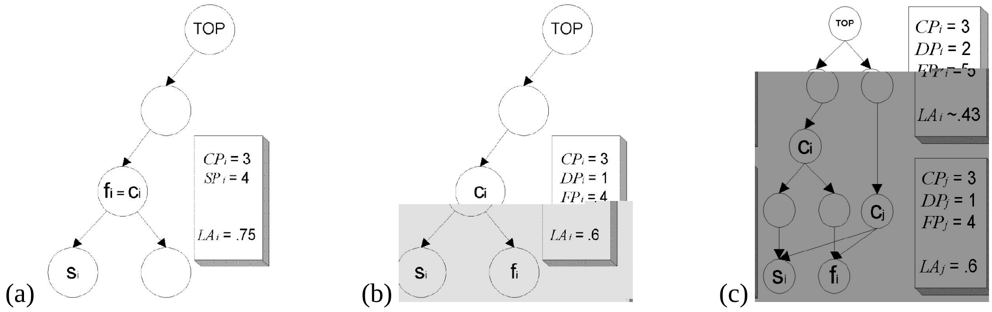


Figure 2: Learning accuracy in three different cases. (a) When the proposed concept is correct, but too general. (b) When the proposed concept is incorrect. (c) When there are different ways to compute Learning Accuracy.

Accuracy calculates the percentage of the proposed hypernyms that are correct:

$$Accuracy = \frac{size(Correct\ hypernyms)}{\sum_i |H_i|} \quad (1)$$

Parsimony is the percentage of concepts for which the set of correct generalisations is equal to the set of suggested generalisations:

$$Parsimony = \frac{|\{u_i : u_i = c_j \quad H_i = G_j\}|}{|U|} \quad (2)$$

Recall is a weaker measure than parsimony. It measures, from the relevant domain-specific concepts (U), the percentage that were correctly identified as relevant and next correctly classified. We say that a concept was correctly classified if at least one of its hypernyms was found.

$$Recall = \frac{|\{u_i : u_i = c_j \quad g \ G_j \text{such that} \ g \ H_i\}|}{|U|} \quad (3)$$

Precision measures, from the chosen concepts, the percentage that were correctly classified in the ontology:

$$Precision = \frac{|\{c_j : c_j = u_i \quad G_j = H_i\}|}{|C|} \quad (4)$$

Production is the mean number of hypothesis generated for each unknown concept.

$$Production = \frac{1}{|C|} \sum_i |H_i| \quad (5)$$

While the first four metrics have to be as high as possible, production is more a descriptive metric. As the other metrics approach 1, production will approach the mean number of hypernyms that the human annotator chose for each domain-specific concept, $\frac{1}{|\mathcal{U}|} \sum_i |G_j|$.

3.4. Distance-based evaluation metrics

When $|H_i| = |G_i| = 1$, it is possible to calculate how large is the distance, in the ontology, between the proposed hypernym and the correct one, using the metric called **Learning Accuracy** (Hahn and Schnattinger, 1998). Let us suppose that the target answer for classifying the unknown concept u_i is s_i, and the system returns the concept f_i. Let us call c_i the lowest concept that is a hypernym of both s_i and f_i. If we call CP_i, SP_i and FP_i the lengths of the shortest paths from the top of the hierarchy to c_i, s_i and f_i, respectively; and DP_i the distance between c_i and f_i, then the Learning Accuracy for u_i is

$$LA_i = \begin{cases} \frac{CP_i}{SP_i} & \text{if } f_i = c_i \\ \frac{CP_i}{FP_i} & \text{if } f_i = s_i \\ \frac{CP_i}{FP_i + DP_i} & \text{otherwise} \end{cases} \quad (6)$$

The overall learning accuracy is the mean of the computed values:

$$LA = \sum_{i \epsilon \{1 \ldots n\}} \frac{LA_i}{n} \quad (7)$$

Figure 2 (a) and (b) show the value of the learning accuracy in two different cases. If the output is correct, Learning Accuracy will have a value of 1.

Because WordNet is not a tree, i.e. a synset can have more than one hypernym, it may be the case that there are several ways to calculate Learning Accuracy, such as that in Figure 2 (c). We have redefined LA as the maximum of all of them, which corresponds to the shortest path between s_i and f_i. Therefore, LA in the example displayed would be 0.6.

However, Learning Accuracy does not take into account that the conceptual distance between a parent node and a chia(bet09(node)et09ian)-289han ontologebet09(ntn)-289(coistact.)433

WordNet we can Is the earth of and

is the ianderns)-255(talt)-769(the)]TJ 148.6076 -11.956 Td[(distancs.)-297(e)15(prre

the first two conceptt are (more t Therefore, the other565(.)]TJ 14.173 -12.729 Td[Usning the

larttcy s_i and of the two conceptt

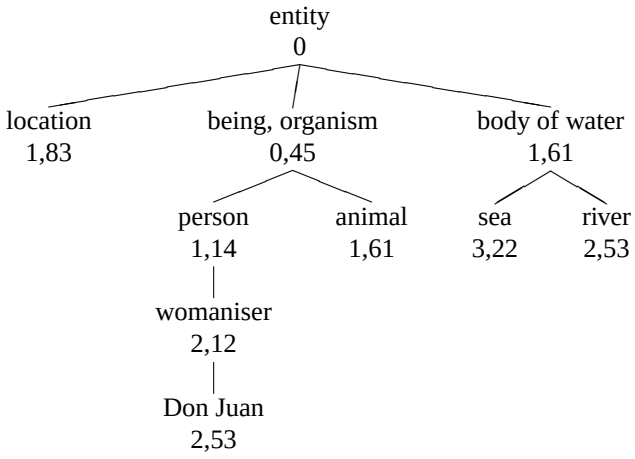


Figure 3: Example of taxonomy in which each node is labelled with its Information Content.

Concept	Freq.	Acc. Freq.	I.C.
entity	0	25	0
location	4	4	1.83
being	3	16	0.45
water	2	5	1.61
person	5	8	1.14
animal	5	5	1.61
womaniser	1	3	2.12
Don Juan	2	2	2.53
sea	1	1	3.22
river	2	2	2.53

Table 2: The concepts in the taxonomy, a hypothetical frequency for each concept, the results of adding up the frequencies of a concept's children, and the Information Content for every concept..

The aim is to maximise ICC_i and to minimise both ICS_i and ICF_i. Therefore, an algorithm has to maximise the following function:

$$ICC_i - ICS_i - ICF_i \quad (8)$$

g_j	ICC_j	ICS_j	ICF_j
animal	1.61	0	0
*womaniser	0.45	1.16	1.67
*Don Juan	0.45	1.16	2.08
*location	0	1.61	1.83

Table 3: Possible generalisations suggested by a classifier, and the values of the three metrics that take into account the Information Content of each node in the ontology.

For example, if we have the ontology in figure 3, and the concepts appear in the ontology with the frequencies shown in table 2, then the Information Content for each concept is the one shown in the figure. Therefore, if we are classifying the new concept *lice*, which should be classified under *animal*, the value of the metrics based on Information Content for several possible outcomes of the classifier is shown in Table 3.

4. Test corpora

As said before, the test corpora has been built from two resources: a portion of *The Wall Street Journal* (WSJ) section in the Penn Treebank, and *The Iliad*. These documents have been pre-processed with the following tools:

- A tokeniser and a sentence-splitter written with regular expressions, in flex.
- The TnT part-of-speech tagger (Brants, 2000).
- A stemmer written in flex.
- Two chunkers written in Java, one for detecting base Noun Phrases, and the other to detect complex verbs. Both use transformation lists (Ramshaw and Marcus, 1995).
- A subject-verb and verb-object detector, written in Java *ad hoc*.

Next, we automatically extracted all the common nouns that were not in WordNet, together with all the sequences of proper nouns. We annotated all of them in the WSJ corpus with the expected hypernyms from WordNet; while in *The Iliad* we only marked the ones with a frequency higher or equal to 50.

These concepts were examined by hand, and classified in some of the following classes:

- A known word with a spelling mistake.
- A previously unknown word. In this case, we identified the WordNet concepts that can be considered its maximally specific generalisations of this word.
- A proper name already in WordNet. In this case, the new concept was annotated with the WordNet synset id.

Figure 4 shows an sample sentence from the corpus, the annotation that it was given and the proposed classification of the unknown concepts and all the proper nouns in the sentence.

5. Conclusions and future work

We have observed that there is strong disagreement about what is included in an Ontology Refinement task, and how to evaluate it. Existing work use different training and test data, ontologies and evaluation metrics. To address this problem, we have built and freely distributed the following framework:

1. A formal definition of the *General Named Entity Identification* task consisting in extending an ontology with new concepts learnt from domain-specific texts. This task can be considered an important subproblem inside OR.
2. Several standard metrics to evaluate it.

```
<s id="396">
    <np det="none" person="3" number="singular" id="397" synsetId="n.wsj.00000033">
        <w c="w" abbreviation="yes" pos="NNP" stem="Dr" id="398">Dr.</w>
        <w c="w" pos="NNP" stem="Talcott" head="yes" id="399">Talcott</w>
    </np>
    <vbar time="past" tense="finite" id="400" subject="397" head="yes" args="+19947">
        <w c="w" pos="VBD" stem="lead" lexhead="yes" head="yes" id="401">led</w>
    </vbar>
    <np id="19947" conjunction="yes">
        <np id="19945" conjunction="yes" head="yes">
            <np det="indefinite" person="3" number="singular" id="402" head="yes">
                <w c="w" pos="DT" id="403">a</w>
                <w c="w" pos="NN" stem="team" head="yes" id="404">team</w>
            </np>
            <pp id="19939">
                <w c="w" pos="IN" id="405" head="yes">of</w>
                <np det="none" person="3" number="plural" id="406">
                    <w c="w" pos="NNS" stem="researcher" head="yes" id="407">researchers</w>
                </np>
            </pp>
            <pp id="19941">
                <w c="w" pos="IN" id="408" head="yes">from</w>
                <np det="definite" person="3" number="singular" id="409">
                    <w c="w" pos="DT" id="410">the</w>
                    <np id="22124" synsetId="n.wsj.00000124">
                        <w c="w" pos="NNP" stem="National" id="411">National</w>
                        <w c="w" pos="NNP" stem="Cancer" id="412">Cancer</w>
                        <w c="w" pos="NNP" stem="Institute" head="yes" id="413">Institute</w>
                    </np>
                </np>
            </pp>
            <w c="w" pos="CC" id="414">and</w>
            <np det="definite" person="3" number="plural" id="415" head="yes">
                <w c="w" pos="DT" id="416">the</w>
                <w c="w" pos="JJ" id="417">medical</w>
                <w c="w" pos="NNS" stem="school" head="yes" id="418">schools</w>
            </np>
            <pp id="19943">
                <w c="w" pos="IN" id="419" head="yes">of</w>
                <np det="none" person="3" number="singular" id="420" synsetId="n.wsj.00000369">
                    <w c="w" pos="NNP" stem="Harvard" id="421">Harvard</w>
                    <w c="w" pos="NNP" stem="University" head="yes" id="422">University</w>
                </np>
            </pp>
        </np>
        <w c="w" pos="CC" id="423">and</w>
        <np det="none" person="3" number="singular" id="424" head="yes" synsetId="n.wsj.00000382">
            <w c="w" pos="NNP" stem="Boston" id="425">Boston</w>
            <w c="w" pos="NNP" stem="University" head="yes" id="426">University</w>
        </np>
    </np>
</s>
```

Synset id	Words	Hypernyms
n.wsj.00000033	James A. Talcott, Dr. Talcott	man, researcher, oncologist
n.wsj.00000124	National Cancer Institute	institute, hospital
n.wsj.00000369	Harvard University	*already in WordNet*
n.wsj.00000382	Boston University	university

Figure 4: Example of sentence annotated. All the processing was done automatically, and we only revised the co-reference of the unknown concepts and annotated the proposed generalisations from WordNet. As can be seen, our automatic parser sometimes fails when parsing conjunctions and when deciding PP-attachment. There are four concepts marked in this sentence, and their annotation is provided in the table.

3. A benchmark test corpus, consisting in financial texts taken from the Wall Street Journal corpus from the Penn Treebank (Marcus et al., 1993) and mythological texts from Homer's *The Iliad*.

This work does not attempt to evaluate learning of non-taxonomic relations (e.g. meronymy, holonymy, telic, etc.), but we believe that similar evaluation metrics could be used (Maedche and Staab, 2000). Further work can be done on this topic.

6. Acknowledgements

This work has been partially sponsored by CICYT, project number TIC2001-0685-C02-01.

7. References

E. Alfonseca and S. Manandhar. 2002. An unsupervised method for general named entity recognition and automated concept discovery. In *Poceedings of the First International Conference on General WordNet*, Mysore, India.

T. Brants. 2000. *TnT - A Statistical Part-of-Speech Tagger*. User manual.

D. Faure and C. Nédellec. 1998. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *LREC workshop on Adapting lexical and corpus resources to sublanguages and applications*, Granada, Spain.

G. Grefenstette. 1993. Automatic thesaurus generation from raw text using knowledge-poor techniques. In *Making Sense of Words. Ninth Annual Conference of the UW Centre for the New OED and text Research*.

U. Hahn and K. Schnattinger. 1998. Towards text knowledge engineering. In *AAAI/IAAI*, pages 524–531.

P. M. Hastings. 1994. *Automatic acquisition of word meaning from context*. University of Michigan, Dissertation.

M. A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING-92*, Nantes, France.

M. A. Hearst, 1998. *Automated Discovery of WordNet Relations. In Christiane Fellbaum (Ed.) WordNet: An Electronic Lexical Database*, pages 132–152. MIT Press.

J. Kietz, A. Maedche, and R. Volz. 2000. A method for semi-automatic ontology acquisition from a corporate intranet. In *Workshop "Ontologies and text", co-located with EKAW'2000*, Juan-les-Pins, French Riviera.

S. Lytinen. 1991. A unification-based, integrated natural language processing system. *Computers and Mathematics with Applications*, 23(6-9):403–418.

A. Maedche and S. Staab. 2000. Discovering conceptual relations from text. In *Technical Report 399, Institute AIFB, Karlsruhe University*.

A. Maedche and S. Staab. 2001. Ontology learning for the semantic web. *IEEE Intelligent systems*, 16(2).

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

L. A. Ramshaw and M. P. Marcus. 1995. Text chunking using transformation-based learning. In *Third ACL Workshop on Very Large Corpora*, pages 82–94. Kluwer.

P. Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis. Department of Computer and Information Science, University of Pennsylvania.

G. Rigau. 1998. *Automatic Acquisition of Lexical Knowledge from MRDs*. PhD Thesis, Departament de Llenguatges i Sistemes Informàtics.– Universitat Politècnica de Catalunya. – Barcelona.

A. Roventini, A. Alonge, F. Bertagna, N. Calzolari, R. Marinelli, B. Magnini, M. Speranza, and A. Zampolli. 2002. In *Poceedings of the First International Conference on General WordNet*, Mysore, India, january.

J. R. R. Tolkien. 1968. *The Lord of the Rings*. Allen and Unwin.

Y. A. Wilks, B. M. Slator, and L. M. Guthrie. 1996. *Electric words: Dictionaries, computers and meanings*. Cambridge, MA: MIT Press.

W. Woods and J. Schmolze. 1992. The kl-one family. *Computer and Mathematics with Applications*, 23(2–5):133–177.