EVALUATING THE HYPONYM ATTACHMENTS IN AN UNSUPERVISED TAXONOMY ENRICHMENT FRAMEWORK

Emil Şt. Chifu and Viorica R. Chifu
Department of Computer Science, Technical University of Cluj-Napoca, Bariţiu 28, Cluj-Napoca, Romania

Keywords: Taxonomy enrichment, unsupervised neural network, extended growing hierarchical self-organizing maps.

Abstract: The paper describes an unsupervised framework for domain taxonomy enrichment with new domain-specific concepts extracted from domain text corpora. The framework is based on an extended model of hierarchical self-organizing maps. Terms extracted by mining a text corpus encode contextual content information, in a distributional vector space. The enrichment behaves like a classification of the extracted terms into the existing taxonomy by attaching them as hyponyms for the intermediate and leaf nodes of the taxonomy. We propose an evaluation setting in which we assess the power of attraction of the population of terms towards the branches of the taxonomy (recall) and the precision of attaching correct hyponyms (accuracy).

1 INTRODUCTION

Our framework for taxonomy enrichment is based on an extended model of hierarchical self-organizing maps, which represent an unsupervised neural network architecture. The candidates for labels of newly inserted concepts are terms collected by mining a text corpus. Each term encodes contextual content information, in a distributional vector space.

Unsupervised hierarchical neural models in general start the growing of a dynamic tree-like topology from a single initial node. Our neural network model, called Enrich-GHSOM, is an extension of one of these existent systems, GHSOM (Dittenbach *et al*., 2002), and it allows the growing to start from an initial tree.

1.1 Related Approaches and Evaluation Strategies

There are two main categories of approaches for taxonomy enrichment (Buitelaar *et al*., 2005): methods based on *distributional similarity* and *clustering of terms*, and approaches using *lexico-syntactic patterns*. Our enrichment approach belongs to the former category, and we will insist in what follows on that category.

In the term clustering approach, the terms extracted from a domain specific corpus of text are classified into an existent taxonomy (Maedche *et al*., 2003; Cimiano and Völker, 2005; Alfonseca and Manandhar, 2002a; Witschel, 2005). In a top-down variant of this classification (Maedche *et al*., 2003; Alfonseca and Manandhar, 2002a; Witschel, 2005), there is a top-down search on the existent taxonomy in order to find a node under which a new term is to be inserted as a successor (hyponym). The classification of the terms is made according to a similarity measure in a distributional vector space. Each term is represented as a vector with information about different contexts of its occurrences in the corpus.

The top-down classification behaviour in our framework is modelled by a growing hierarchical self-organizing map (GHSOM) architecture (Dittenbach *et al*., 2002) extended with the possibility to set an initial state for the tree-like neural network. In our new extended neural model, called Enrich-GHSOM, the existing taxonomy is given as the initial state of the neural network. The model allows to classify the extracted terms into the existing taxonomy by attaching them as hyponyms for the intermediate and leaf nodes of the taxonomy. Details of this process are given in section 3.

(Widdows, 2003) is searching for a node to attach a new concept as a hyponym, by finding a place in the existent taxonomy where the corpus derived semantic neighbours of the candidate concept are most concentrated. He supposes that at

Şt. Chifu E. and R. Chifu V. (2008).
EVALUATING THE HYPONYM ATTACHMENTS IN AN UNSUPERVISED TAXONOMY ENRICHMENT FRAMEWORK.
In *Proceedings of the Fourth International Conference on Web Information Systems and Technologies*, pages 247-252
DOI: 10.5220/0001529302470252

least some of the semantic neighbours are already in the taxonomy.

(Alfonseca and Manandhar, 2002a) evaluate their framework by measuring the *strict* and *lenient accuracy* of attaching new concepts as hyponyms. The lenient accuracy assesses more indulgently as correct a classification of a new concept as a hyponym of an existent concept. They also assess the *learning accuracy*, by taking into account the distance in the taxonomy between the chosen hypernym (mother) node of the candidate and the correct one. In the learning accuracy, incorrectly classified new concepts are given a weight inverse proportional to this taxonomic distance, instead of counting as zero, like in the strict accuracy evaluation. In this sense, learning accuracy can be considered as a kind of lenient accuracy. All of these evaluations can be assessed by a human expert in the domain which is able to say whether a pair of concepts is in the hyponym-hypernym (is-a) relation.

(Witschel, 2005) also assesses his framework by using accuracy and learning accuracy. To be independent from a human expert, he evaluates the correct hyponym attachments by actually using subtrees from GermaNet (German WordNet) and only proposing "new" concepts which are in GermaNet (but provisionally removed before their experimental reinsertion). (Witschel, 2005) also proposes a qualitative analysis experiment which classifies terms into two main taxonomies. The terms to classify are known (from GermaNet) to belong semantically either to one tree or to the other tree. The discrimination power of his decision tree based classification approach is assessed by computing a *recall* and *precision* of classifying the terms of the two categories into the two main trees. Out of the total number of terms in a given category, he defines a recall as the percentage of terms of that category which will get classified into either one of the trees, no matter which of them (the rest being not attracted by any of the two proposed trees), out of which the precision represents the percentage of terms classified into the correct tree. We will evaluate in section 4.1 this kind of recall (Witschel, 2005), even if it is not the *standard recall* measure, which rather counts only the *correctly* classified terms out of the total number of terms in a category.

Like (Witschel, 2005), (Widdows, 2003) tries to reconstruct WordNet (Fellbaum, 1998) and thus he is independent from a human expert. He measures the accuracy and a lenient/learning accuracy, the latter being accounted for as starting from the number taxonomic levels (as measured in WordNet) between the chosen existent hypernym and the new correct hyponym. He thus considers correct classifications of new concepts with different levels of detail. For instance, the new concept *cat* can be attached as hyponym under the concept *feline*, *carnivore*, *mammal* or *animal* with different levels of detail as a consequence of different hypernym-hyponym taxonomic distances.

2 LEARNING TECHNIQUE

Our model of hierarchical self-organizing maps – Enrich-GHSOM – represents the unsupervised neural network based learning solution adopted by our taxonomy enrichment framework. This choice fits well with the knowledge structure to be enriched – a taxonomy, i.e. an *is-a* hierarchy of concepts. The neural model is an adaptation of GHSOM, an existent unsupervised neural system (Dittenbach *et al*., 2002).

GHSOM is an extension of the Self-Organizing Map (SOM) learning architecture (Kohonen *et al*., 2000). A thesaurus is a data space consisting of terms in a language, represented as a lexical data base. The main relation between the terms in a thesaurus is the taxonomic relation. However, because of their essentially flat topology, SOM maps have a limited capability to discover and illustrate hierarchical clusters in data sets. The growing hierarchical self-organizing map model consists of a set of SOM maps arranged as nodes in a hierarchy and it is able to discover hierarchical clusters (Dittenbach *et al*., 2002).

2.1 Enrich-GHSOM

The growth of a GHSOM is a completely unsupervised process, being only driven by the unlabeled input data items themselves together with the two thresholds and some additional learning parameters. There is no way to suggest from outside any initial paths for the final learnt hierarchy. We have extended the GHSOM model with the possibility to force the growth of the hierarchy along with some predefined paths of a given initial hierarchy. Our new extended model, Enrich-GHSOM, is doing a classification of the data items into an existing taxonomic structure. This initial tree plays the role of an initial state for the tree-like neural network model. The classical GHSOM model can only grow as starting from a single node. The top-down growth in our extended model starts from a given initial tree and inserts new nodes attached as successors to any of its intermediate and leaf nodes.

More details of this process are given in section 3, in the specific context of taxonomy enrichment, when we classify terms extracted from a corpus into a given taxonomy.

3 A FRAMEWORK FOR UNSUPERVISED TAXONOMY ENRICHMENT

The whole processing in our framework can be divided into two main steps: the *term extraction* step and the *taxonomy enrichment* step. The candidates for labels of new concepts inserted during the taxonomy enrichment are terms identified by mining the domain text corpus. Three different linguistic entities can be recognized as terms in the current framework: simply words, nouns, and noun phrases. In order to identify such term categories by a linguistic analysis of the corpus documents, our framework relies on several processing resources offered by the GATE framework (Cunningham *et al*., 2002).

By reference to the classification in (Buitelaar *et al*., 2005), our approach to taxonomy enrichment is one based on *distributional similarity and term clustering*. The terms extracted from the text corpus are mapped into the existing taxonomy, which plays the role of an initial skeleton class system to suggest term classification. The taxonomy enrichment algorithm proceeds by classifying the terms collected from the corpus into the given taxonomy. The Enrich-GHSOM neural network drives a top-down hierarchical clustering of the terms along with the given taxonomy branches and inserts new nodes (concepts) corresponding to these classified terms. Every new concept is attached as successor to either an intermediate or a leaf node of the given taxonomy and becomes a hyponym of that node.

In order to use our Enrich-GHSOM neural network to induce such a taxonomy enrichment behaviour, a symbolic-neural translation is first done by parsing a textual representation of the initial taxonomy (*is_a*(*concept*, *superconcept*) assertions or OWL format). The result of this parsing is the initial internal tree-like state of the neural network. In order for the initialized network to be able to classify terms into this initial taxonomic structure, a representation as a numerical vector is needed for each node in the initial taxonomy. This will be the vector representation for the concept label associated to the node, computed as described in section 3.1. We assume that the concept labels of the initial taxonomy are terms extractable from the domain text corpus used in the taxonomy enrichment. Their vectors are then computed in the same way as the vectors of all the corpus extracted terms which will be classified during the enrichment. Using the same corpus from a specialized domain to acquire the feature vectors of the concepts in the initial taxonomy and the terms to be classified is a reasonable choice, since it will reduce the problems with the ambiguous term senses.

3.1 Term Vector Representation

In our framework, the context features of the vector representation of a term are the frequencies of the term occurrence in different documents of the corpus. The Euclidean distance is used to compute similarity among term vectors.

The framework allows multiple ways to encode the frequencies of term occurrence, starting from simple *flat counts of occurrences*. Another variant is *DF-ITF* (document frequency times inverse term frequency) (Chifu and Leția, 2006). A third way to encode the term vector representation is one in which we propose the vector to be a *document category histogram* (Chifu and Leția, 2006). The *dimensionality reduction* achieved by this representation is important since it removes the semantic noise caused by minor differences in the semantic content of different documents in the corpus. This intuition is confirmed by the experiments described in section 4. Moreover, the term/document occurrence matrix is sparse and a more natural behaviour of the neural model is expected by using reduced and less sparse vectors.

3.2 Data Sparseness and Average Vector

We just ended the last subsection with a conclusion that sparse vectors should be avoided in our framework by reducing their dimensionality with the help of the document category histograms. A source of data sparseness is represented by *terms with very few occurrences* in the text corpus. This is the case of the most generic terms which label the roots of the main trees in a given initial taxonomy. When the Enrich-GHSOM neural network is given a very sparse vector of such an overly generic concept as one of the roots, then the main tree rooted by that concept is unable to attract and classify a relevant quantity of terms.

(Alfonseca and Manandhar, 2002b) and (Witschel, 2005) acknowledged the same problem

and proposed to associate to every concept in the initial taxonomy either the sum of the vector representation of the concept itself and its immediate successors (Alfonseca and Manandhar, 2002b), or the sum of the vector representation of all the concepts of the tree rooted by the given concept, including the root itself (Witschel, 2005). We tried the second approach (Witschel, 2005) without any improvement noticed. However, taking the *average vector*, i.e. the *centroid*, instead of the sum (like in (Pekar and Staab, 2002; Maedche *et al*., 2003; Cimiano and Völker, 2005)) has led us to a significant improvement of the experimental results. A similar improvement obtained by using the centroid vector representation for concepts is reported in (Cimiano and Völker, 2005).

4 EXPERIMENTS AND EVALUATION

The experiments carried out in what follows are in the "4 universities" domain (Craven *et al*., 2000). The corpus contains 8185 Web pages from Computer Science departments. We propose an evaluation setting in which we assess the power of attraction of the population of terms towards the branches of the existent taxonomy (*recall*) and the precision of attaching correct hyponyms (*accuracy*).

4.1 Evaluating the Recall (Attraction of Terms)

In this first set of experiments we evaluate the attraction of terms towards the main trees of an initial rather shallow taxonomy, taken exactly from (Craven *et al*., 2000):

entity
  activity
    course
    research_project
  person
    faculty
    staff
    student
  department

We are only interested in assessing the quantity of terms attracted towards the three main trees rooted *activity*, *person*, and *department*. This is in lines with the qualitative experiment in (Witschel, 2005), which measured the discrimination power of his decision tree approach (see section 1.1). Like in (Witschel, 2005), only the choice of one of the main trees was assessed, disregarding any further top-down decisions *within* any of the main trees. That is why a shallow initial taxonomy is enough for the experiment.

We evaluate the quantity of terms classified into the three main trees of our initial taxonomy by using a recall measure as defined in (Witschel, 2005). In contrast to (Witschel, 2005), who computes one recall for each semantic category of terms, we only compute a single *overall recall*, as we don't have any a priori (WordNet based like (Witschel, 2005)) semantic categorization of the terms extracted by mining the corpus. Also many of our extracted terms maybe don't belong semantically to any of our three main taxonomies. Out of the total number of terms extracted, the overall recall computes the percentage of terms which get classified into our three main taxonomies rooted *activity*, *person*, and *department*, no matter which of them (the rest being not attracted by any of the three proposed trees, being rather organized in self-defined new trees). Table 1 shows the results of four experiments which differ in the chosen settings for the term linguistic category and term vector representation.

Table 1: Attraction of terms towards the main trees activity, person and department.

Experiment	1	2	3	4
Linguistic Category	noun	noun phrase	noun phrase	word
Vector Represent.	histogram	Df-Itf	Df-Itf	flat counts
activity	83	6	15	28
person	78	20	9	39
department	39	---	---	7
Total # of terms extracted	887	2004	2004	2798
Overall Recall	22.6%	1.3%	1.2%	2.7%

The initial taxonomy contains the noun phrase *research project*, which is represented simply as *project* in experiments 1, 3 and 4. We decided to use *project* instead of *research project*, as we were constrained to use a noun and a word respectively as concept identifiers in the initial taxonomy in experiments 1 and 4 (see section 3). The difference between experiments 2 and 3, both having noun phrases as terms, is that in 2 we actually set the concept in the initial taxonomy as the noun phrase *research project* as opposed to the noun phrase *project* in 3. Also in cases 2 and 3, the only occurrences of the noun *department* in the corpus

were as aggregated in compound noun phrases which are more specific than the singleton noun phrase *department*. That is why in experiments 2 and 3 we don't have any tree rooted *department*.

Finally, it turns out that the best results are with the histogram representation of terms. This confirms the expectations in sections 3.1 and 3.2 about dimensionality reduction and data sparseness. In Table 1 and formula (1), *activity*, *person* and *department* represent the number of terms classified into the main trees rooted *activity*, *person*, and *department* of the given initial taxonomy, and formula (1) computes the overall recall.

$$\text{Overall Recall} = \frac{activity + person + department}{\text{Total \# of terms extracted}} \quad (1)$$

4.2 Evaluating the Accuracy of Attaching Correct Hyponyms

Now we turn to the question of measuring the classification accuracy of the enrichment, i.e. how many of the terms classified into the different main trees of an existing taxonomy are inserted as correct hyponyms. We actually use a *lenient accuracy* measure similar in spirit to the one evaluated in (Alfonseca and Manandhar, 2002a) and (Hearst and Schütze, 1993).

Since we now assess the accuracy of the attached new concepts as hyponyms to existing nodes, we need a deeper initial taxonomy (with an increased number of existing nodes). The new taxonomy, shown in Figure 1, originates from the above "4 universities" taxonomy, combined with some knowledge from the corresponding domain specific WordNet branches.

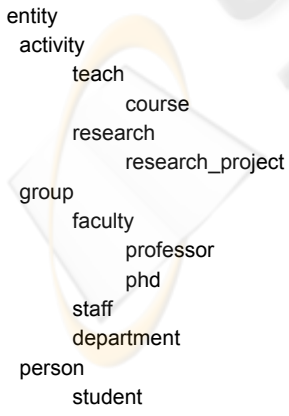

Figure 1: Initial taxonomy used for evaluating the accuracy of inserting new hyponyms.

Table 2: Accuracy of attaching new hyponyms.

Experiment	1	2
Linguistic Category	noun	word
Vector Representation	histograms	flat counts
Total # of attachments	166	87
Total # of correct hyponym attachments	92	69
Lenient Accuracy	55.4%	79.3%

Table 2 shows the *lenient accuracy* results of two experiments having the same settings like experiments 1 and 4 in section 4.1, for which better recall has been obtained. *Total # of attachments* is the number of new nodes inserted as successors to intermediate and leaf nodes of the existent taxonomy. Out of this quantity, *Total # of correct hyponym attachments* is the number of inserted nodes which can be considered as valid hyponyms of the nodes to which they are attached as successors. Formula (2) computes the lenient accuracy.

$$\text{Lenient Accuracy} = \frac{\text{Total \# of correct hyponym attach.}}{\text{Total \# of attachments}} \quad (2)$$

Table 3: Correctly inserted new nodes as successors of three nodes of the initial taxonomy in Figure 1.

Extant Node	Correct Hyponym Attachments
faculty	technology, engineer, control, center, compute
person	graduate, study, learn, government
course	class, hour, syllabus, assignment, fall, oct, tuesday

Table 3 gives a couple of example new nodes inserted by our framework, which we considered as correct hyponym attachments for evaluating the lenient accuracy of enriching the initial taxonomy in Figure 1. The new nodes are assigned to the correct topic (*faculty*, *person*, *course* respectively) even if not all of them are strict hyponyms of one of the three concepts. Two of the existent concepts are intermediate nodes, and one is a leaf of the initial taxonomy. For an intermediate node, the newly inserted concepts are brothers of their extant hyponyms. For instance, *[faculty of] technology*, *[faculty of] engineering*, *center* etc. become brothers with *professor* and *phd*, even if the latter rather correspond to the North American meaning of the concept of *faculty*, as faculty member.

Finally, it can be noticed from Table 2 that the histogram vector representation for terms gives lower lenient accuracy as compared to the flat counts. But remember from the previous subsection

that the recall is much better for the histogram representation. A better recall is indeed expected to come together with a reduced precision (accuracy). In general in ontology learning the recall is more important than precision. For a domain expert it is better to delete many spurious (wrong) hyponym attachments than to miss other correct attachments because of a reduced recall.

5 CONCLUSIONS AND FURTHER WORK

We have presented an unsupervised top-down neural network based approach and framework for taxonomy enrichment. The experimental results obtained in the "4 universities" domain are encouraging, especially when the terms extracted from the corpus are represented with a reduced dimensionality, as *document category histograms*. Moreover, the best enrichment results have been achieved when we chose the *average vector* to represent every concept in the given initial taxonomy.

Our framework can also be used as a tool to assist a domain expert in building an ontology. From this point of view, the recall is more important than the accuracy. The domain expert has to manually prune out the wrong hyponym attachments. As further work to better evaluate our taxonomy enrichment framework we will rely more on existing thesauri like WordNet for evaluating the quality of the hyponym attachments.

ACKNOWLEDGEMENTS

This work was supported by the grant TD-416 (520/2007) from the National Research Council of the Romanian Ministry of Education and Research.

REFERENCES

Alfonseca, E., Manandhar, S., 2002. Extending a lexical ontology by a combination of distributional semantics signatures. In A. Gómez-Pérez, V.R. Benjamins (Eds.), *13th International Conference on Knowledge Engineering and Knowledge Management, LNAI*. Springer, pp. 1-7.

Alfonseca E., Manandhar, S., 2002. An unsupervised method for general named entity recognition and automated concept discovery. In *1st International Conference on General WordNet*.

Buitelaar, P., Cimiano, P., Magnini B., 2005. Ontology learning from text: an overview. In P. Buitelaar, P. Cimiano, B. Magnini (Eds.), *Ontology Learning from Text: Methods, Evaluation and Applications*, Frontiers in Artificial Intelligence and Applications Series. IOS Press, pp. 1-10.

Chifu, E.Ş., Leţia, I.A, 2006. Unsupervised ontology enrichment with hierarchical self-organizing maps. In I.A. Leţia (Ed.), *IEEE 2nd International Conference on Intelligent Computer Communication and Processing*, pp. 3-9.

Cimiano, P., Völker, J., 2005. Towards large-scale, open-domain and ontology-based named entity classification. In *RANLP'05, International Conference on Recent Advances in Natural Language Processing*, pp. 166-172.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., Slattery, S., 2000. Learning to construct knowledge bases from the World Wide Web. *Artificial Intelligence* 118, pp. 69-113.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., 2002. GATE: a framework and graphical development environment for robust NLP tools and applications. In *40th Anniversary Meeting of the ACL*.

Dittenbach, M., Merkl, D., Rauber, A., 2002. Organizing and exploring high-dimensional data with the Growing Hierarchical Self-Organizing Map. In L. Wang, *et al*. (Eds.), *1st International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 2, pp. 626-630.

Fellbaum, Chr. (Ed.), 1998. *WordNet: An Electronic Lexical Database*, MIT Press. Cambridge, Mass.

Hearst, M.A., Schütze, H., 1993. Customizing a lexicon to better suit a computational task. In *ACL SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*, 1993.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., Saarela, A., 2000. Self-organization of a massive document collection. *IEEE Transactions on Neural Networks* 11, pp. 574-585.

Maedche, A., Pekar, V., Staab, S., 2003. Ontology learning part one: on discovering taxonomic relations from the Web. In N. Zhong, *et al*. (Eds.), *Web Intelligence, LNCS*. Springer, pp. 301-321.

Widdows, D., 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *HLT-NAACL Conference*, pp. 197-204.

Witschel, H.F., 2005. Using decision trees and text mining techniques for extending taxonomies. In *Learning and Extending Lexical Ontologies by using Machine Learning Methods, Workshop at ICML-05*, pp. 61-68.