Building phylogenetic lexical ontologies*

Enrique Alfonseca
Computer Science Department
Universidad Autonoma de Madrid
Enrique.Alfonseca@ii.uam.es

Abstract

This paper describes a new application of the theory of cladistics to automatic acquisition of lexical ontologies. In the case of life forms, those which are phylogenetically related are expected to share common properties and hence to appear in similar contexts in texts. The procedure has been tested with other words that do not refer to life beings, and a standard phylogenetic algorithm is used to learn automatically lexical ontologies from free text. The resulting ontologies are semantically coherent, as shown by comparing them with WordNet.

1 Introduction

Cladistics is a method for phylogenetic classification initially proposed by the German entomologist W. Hennig. Compared to previous approaches to phylogeny, it offered the advantage that it was based on a repeatable method of evaluating relationships. Cladistics studies the systematic classification of groups of organisms based on the characteristics that are shared among them.

The method of cladistics is based on three basic assumptions. The first of them states that any group of organisms has a common antecessor. Secondly, it is assumed that the taxonomic trees will be binary. Finally, evolutionary changes (mutations) occur over time. Some approaches might make further assumptions. Finally, the method for scoring trees, which assigns the highest score to a tree with the least amount of mutations, is called *parsimony*. For an introduction to cladistics, refer to [Page and Holmes, 1998].

The cladistic theories have been applied to study the the evolution of cultures [Alfonseca, 1979; Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985], anthropology [Holden and Mace, 2002], and language evolution [Gray and Jordan, 2000].

1.1 A phylogenetic theory of lexical semantics

It might come as a surprise that lexical ontologies of life forms tend to parallel biological phylogenetic trees. In most examples of lexical ontologies that contain living organisms, these have been structured in the same way as they would be by a biologist. This can be observed in general-purpose lexical ontologies such as WordNet [Miller, 1995] or the Cyc upper level ontology [Lenat and Guha, 1990]. However, in reality, life beings which are very close in the evolutionary tree usually share many properties, such as their appearance and abilities, and therefore they will appear near a similar set of words. For instance, most names for mammals can appear with the adjective *furred* or with the noun *placenta*, and only subtypes of horses can appear as subject of the verb *neigh* or *whinny*). More specialised beings are expected to appear in more specialised contexts. In fact, most approaches for automatic ontology construction from text use contextual information to structure the concepts in a tree (e.g. [Lee, 1997; Alfonseca and Manandhar, 2002; Pekar and Staab, 2003]).

Therefore, by examining distributional properties of words it should be possible to generate the most parsimonious tree. For instance, features such as the ability *to ingest*, *to reproduce* or *to breath* can be used by cladistic algorithms to group together most of the organisms that have those faculties as descendants of a single antecessor. If the words *horse*, *roan* and *mare* have those properties, together with a few others, e.g. *to neigh*, *to whinny* and *to gallop*, then they might be grouped together in a cluster.

Furthermore, in principle it might be possible to apply the same argumentation to other kinds of entities apart from life forms. For instance, all kinds of knives can appear in all the contexts in which the term *knife* appears, and possibly in a few ones, more specialised. In this way, in the field of lexical semantics, we can study the progressive specification of the qualities of the concepts represented by words, and use that for obtaining the equivalent to an *evolutionary tree*. The application to ontology building would be in this way: we can characterise every term with a phenotype describing the contexts in which that term appears, and use a cladistic analysis to obtain a phylogenetic tree.

For illustration, let us consider the seven concepts *body of water*, *river*, *organism*, *plant*, *animal*, *bear* and *horse*, and the eight features (eight verbs) listed at the bottom of Figure 1. The Figure also shows the concepts organised in a standard taxonomy, where the concepts are ordered from the most general to the most specific. For each concept and feature, we can assign it the value 1 if the concept has that feature (i.e., if the concept appears in the context of the word that represents the feature), and the value 0 otherwise. The phenotype of each concept is shown below its name as a boolean vector.

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

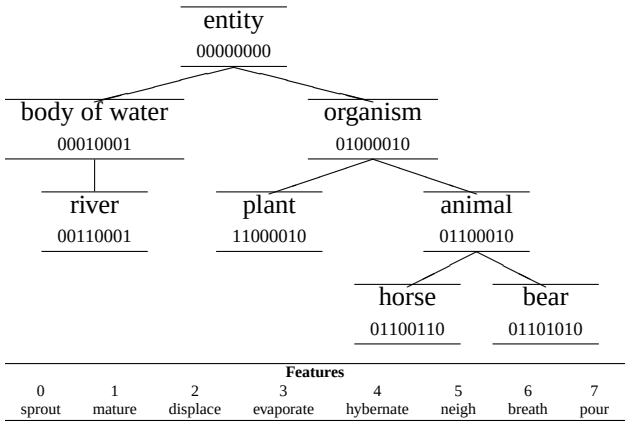

Figure 1: Example ontology of concepts, and features.

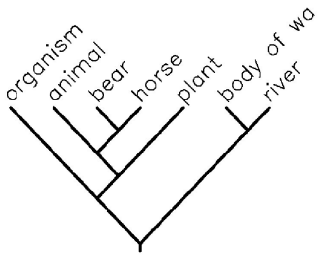

Figure 2: One of the most parsimonious tree obtained with a cladistic analysis of the seven concepts studied.

If we consider the vector of features as the phenotype of each concept, and we establish that the root ancestor is the one for which all features have value 0, then we can use one of the many cladistic algorithms to derive the phylogenetic tree. Figure 2 shows a tree obtained with the *phylip* toolset [Felsenstein, 1993], using the *Dollo* algorithm. As can be seen, it shares the same structure as the conventional tree, except that, in cladistic trees, every concept must always be a leaf node; therefore, instead of placing *organism* as parent of *animal*, it is set as an "uncle" node in the tree.

2 Experiment and conclusions

This method has been used to organise 100 concepts (synsets) chosen randomly from the hierarchy of physical entities in WordNet. The procedure has been the following:

1. For each concept c,
   (a) Download up to 500 documents from the Internet where that concept appears, using the automatic procedure described in [Agirre *et al.*, 2000].
   (b) Parse all the sentences from the documents where any of c's words appears [Alfonseca, 2003].
   (c) Collect automatically all the verbs for which any of c's words was the head of the subject.
   (d) Represent the concept as a boolean vector, where a 1 in the i^{th} position means that c appears as subject of the i^{th} verb.
2. Derive a phylogenetic tree from that data using the *Dollo* procedure [Farris, 1977].

For evaluation, we have pruned WordNet to include only the 100 synsets selected, and we compare it against the phylogenetic tree obtained. The metric used is the following: for each cluster in the original tree, we calculate its F-score compared against the original WordNet [Zhao and Karypis, 2002]. In this first experiment, the mean of the F-scores for every cluster was 54.96%, a result similar to other experiments (not really comparable, as the datasets used are different) [Cimiano *et al.*, 2004].

For future work, we plan to extend this experiment along the following lines: (a) study how the choice of the parsimony algorithm affects the results; (b) extend the description of the phenotypes with adjectives and with words that hold other syntactic dependences; (c) perform a statistical analysis, such as the one we did in [Alfonseca and Manandhar, 2002], to remove from the phenotypes contextual terms that might be due to irregular use of words and very general term which are probably not very informative.

References

[Agirre *et al.*, 2000] E. Agirre, O. Ansa, E. Hovy, and D. Martinez. Enriching very large ontologies using the www. In *Ontology Learning Workshop, ECAI*, Berlin, Germany, 2000.

[Alfonseca and Manandhar, 2002] E. Alfonseca and S. Manandhar. Extending a lexical ontology by a combination of distributional semantics signatures. volume 2473 of *Lecture Notes in Artificial Intelligence*, pages 1–7. Springer Verlag, 2002.

[Alfonseca, 1979] M. Alfonseca. *Human Cultures and Evolution*. Vantage Press, New York, 1979.

[Alfonseca, 2003] E. Alfonseca. Wraetlic user guide v. 1.0, 2003.

[Boyd and Richerson, 1985] R. Boyd and P. Richerson. *Culture and the Evolutionary Process*. Univ. of Chicago Press, 1985.

[Cavalli-Sforza and Feldman, 1981] L. L. Cavalli-Sforza and M. W. Feldman. *Cultural transmission and evolution: a quantitative approach*. Princeton University Press, Princeton, 1981.

[Cimiano *et al.*, 2004] P. Cimiano, A. Hotho, and S. Staab. Clustering concept hierarchies from text. In *Proceedings of LREC-2004*, 2004.

[Farris, 1977] R. J. Farris. Phylogenetic analysis under dollo's law. *Syst Zool*, 26:77–88, 1977.

[Felsenstein, 1993] J. Felsenstein. PHYLIP (Phylogeny Inference Package) version 3.5c. distributed by the author, 1993.

[Gray and Jordan, 2000] R. Gray and F. Jordan. Language trees support the express-train sequence of austronesian expansion. *Nature*, 405:1052–1055, 2000.

[Holden and Mace, 2002] C. Holden and R. Mace. *Pastoralism and the evolution of lactase persistence*, pages 280–307. Cambridge University Press, Cambridge, 2002.

[Lee, 1997] L. Lee. *Similarity-Based Approaches to Natural Language Processing*. Ph.D. thesis. Harvard Univ., 1997.

[Lenat and Guha, 1990] D. B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems*. Addison-Wesley, 1990.

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

[Page and Holmes, 1998] R. Page and E. Holmes. *Molecular Evolution: A Phylogenetic Approach*. Blackwell Pub., 1998.

[Pekar and Staab, 2003] V. Pekar and S. Staab. Word classification based on combined measures of distributional and semantic similarity. In *Research Notes of EACL-03*, 2003.

[Zhao and Karypis, 2002] Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *11th Int. Conf. on Inform. and Knowledge Management*, 2002.