ECAI 2008
M. Ghallab et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-891-5-288

Multilingual Evidence Improves Clustering-based Taxonomy Extraction

Hans Hjelm and **Paul Buitelaar** [1]

Abstract. We present a system for taxonomy extraction, aimed at providing a taxonomic backbone in an ontology learning environment. We follow previous research in using hierarchical clustering based on distributional similarity of the terms in texts. We show that basing the clustering on a comparable corpus in four languages gives a considerable improvement in accuracy compared to using only the monolingual English texts. We also show that hierarchical k-means clustering increases the similarity to the original taxonomy, when compared with a bottom-up agglomerative clustering approach.

1 Introduction

Does a country and its environment form the language of the people living in it? Or does the language spoken rather form the way people perceive their environment? This type of questions has been raised by linguists like Sapir/Whorf and Berlin/Kay. Whatever the answer to such questions, we *do* believe that each language provides us with a unique "view" of the world, coded into its grammar and lexicon. The question we wish to answer in this paper is whether this diversity will prove an asset in a taxonomy extraction system or whether the different "views" will merely serve to clutter the meaning expressed through an isolated language.

Several researchers have made use of clustering based on distributional similarity between terms to perform taxonomy extraction [1, 2, 12, 13]. We follow their work by first extracting a taxonomy using only English language texts and comparing the result to a *gold standard* taxonomy. We then repeat the procedure, building a taxonomy using four different languages (adding German, French and Spanish to the English), using a comparable corpus. We show that the multilingual version gives a considerable improvement in accuracy and stability over the monolingual version, when compared to the gold standard.

We also make use of a hierarchical k-means clustering technique and show that we are able to reproduce the original taxonomy with greater fidelity than when using a bottom-up agglomerative clustering approach.

2 Background and resources

As part of the recently started THESEUS MEDICO project,[2] funded by the German government, a system for querying and analyzing medical information (medical records, x-rays etc.) is currently under construction. Certain parts of such a system would arguably benefit from a domain ontology, providing background knowledge during e.g., information retrieval or image recognition tasks. In the domain of (human) anatomy, there exists such an ontology: the Foundational Model of Anatomy (FMA) ontology. It is developed by the Structural Informatics Group at the University of Washington and it is open source.[3] It contains about 100,000 English terms, 8,000 Latin, 4,000 French, 500 Spanish and 300 German terms. There are also some terms in other languages such as Italian and Filipino, but they were not used in this project. The languages we decided to work with (based on available resources and language competence in the project) were English, German, French and Spanish. The ontology models the hierarchical *is-a* and *part-of* relations, along with some other relations, but only the is-a structure was considered in this project. We define our task as such: given a domain-specific corpus, we want to recreate the structure of the FMA ontology as closely as possible, using a hierarchical term clustering approach.

2.1 Corpus collection

We needed a domain corpus in the relevant languages to have data on which to train the distributional models. We decided to use the Wikipedia[4] pages filed under the 'Anatomy' category for each language.[5] This resulted in about 7,300 pages for English, 2,600 for French, 2,400 for German and 1,000 for Spanish. This corresponds to about 4.4 million words for English, 1.1 million for French, 890,000 for German and 400,000 for Spanish. We stripped the texts of HTML and other markup or scripts, as well as Wikipedia related text (as far as possible). It should be noted that Wikipedia is constantly changing and growing and that these numbers reflect the status as of February 2007.

2.2 Preprocessing the data

In order to lessen some of the detrimental effects of the data sparseness problem, we decided to lemmatize the corpus (giving us more occurrences of each word type). We used Intrafind's[6] LiSa system for morphological analysis[9] for all languages.

Since the concepts in the ontology are associated with *terms* rather than words, we needed a way of letting the automatic methods treat multi-word terms as single units, as well as being able to distinguish single word terms from "mere" words. We therefore made use of a simple term spotting technique (see [10] for more on term spotting), marking the longest consecutive string of words that also appears in

[1] GSLT/Stockholm University, Sweden, email: hans.hjelm@ling.su.se and DFKI, Germany, email: paulb@dfki.de
[2] http://theseus-programm.de/scenarios/en/medico
[3] http://sig.biostr.washington.edu/projects/fm/index.html
[4] http://www.wikipedia.org
[5] For English: http://en.wikipedia.org/wiki/Category:Anatomy. This page links to the corresponding pages in the other languages.
[6] http://www.intrafind.de

the FMA ontology, as a term. The terms in the FMA ontology were also lemmatized, in order to better match the lemmatized corpus text.

After preprocessing the data, it looks something like this (example from the FMA corpus):

```
the TERM_zygomatic_bone#55158 (
malar TERM_bone#34122 ) be a pair
TERM_bone#34122 of the human
TERM_skull#49338 .
```

Obviously many other relations (e.g., the synonymy and part-of relations) also hold between the terms we are studying, not just the is-a relation. The distributional model we apply in these experiments (see Sect. 2.3) is not designed to separate between different types of relatedness – we accept a degree of overlap between relations in our learned taxonomy. However, given the results presented in the papers listed in Sect. 1, we can expect a *focus* on the type of relations we are interested in for these experiments (i.e., the sibling (cohyponymy) and is-a (hyponymy) relations).

2.3 Distributional model parameters

When building the distributional models for each language, there are a number of parameters that can be varied. In a pre-study, we examined the effects of these parameters on a flat clustering task, i.e., merely clustering the terms into groups, with no hierarchical information. We used the best settings from this pre-study for the hierarchical clustering experiments.

Although using the settings from such a pre-study will not be possible in a typical application scenario, our aim here is not to present high-scoring evaluation figures for the system as such. Our focus is to investigate a possible improvement when using multilingual data as opposed to just the monolingual data. Also note that the pre-study was carried out on strictly monolingual data.

Size of sliding window: When constructing a term-term distributional model, one typically makes use of a fixed-size sliding window which is moved over the text. Varying the size of this window effects the type of information captured by the model. We varied the window size between 3–500 in our experiments (on each side of the focus word). We also investigated the effects of *not* using a sliding window, but rather using document co-occurrence as our features.

Minimum feature frequency: If a feature is too infrequent, it is possible that its distribution is not captured well enough in the corpus to be of any use. We therefore experimented with different lower thresholds for our features.

Left/right distinction: In some cases it might be important to keep track of whether the context word appeared to the left or to the right of the focus word. If we want to make this distinction, we simply introduce separate features for each word: one for the left and one for the right context.

Distance weighting: Intuitively, words appearing closer to the focus word should be given more weight than words appearing further away when building a distributional model. We made use of three different distance weighting schemes in our experiments (d stands for distance measured in number of words from the focus word):

1. Flat: no weighting scheme is applied.
2. Inverse distance: the context term is weighted by $\frac{1}{d}$.
3. Logarithmic distance: the context term is weighted by 2^{1-d} (weights decrease faster than for Inverse distance).

Feature weighting: We can hypothesize that a very frequent context word (measured over the whole corpus) contributes less to defining the "distributional profile" of a focus word, than a less frequent context word would. We use three feature weighting schemes to try to model this hypothesis (the choice of weighting schemes is inspired by [2]):

1. Flat: no feature weighting is applied.
2. Conditional probability: if the term under consideration is t, the current feature is f and $freq$ stands for the frequency of a particular term or term-feature pair, then we get:
$$weight(t, f) = p(t|f) \approx \frac{freq(t, f)}{freq(f)}$$
3. Mutual Information: we can write the Mutual Information formula like this:
$$\sum_{t_x, f_y} p(t_x, f_y) log \frac{p(t_x, f_y)}{p(t_x)p(f_y)}$$
where $x, y \in \{0, 1\}$, indicating the presence or absence of t and f (again, probabilities are estimated using relative frequencies).

Dimensionality reduction: we tried using singular value decomposition [6] for some settings on the distributional models.

3 Hierarchical clustering

We examine two kinds of hierarchical clustering: bottom-up agglomerative clustering and hierarchical k-means. Neither method produces a hierarchy in the traditional sense, but rather a structure like the one depicted in figure 1. The bottom-up approach builds this structure starting with each term in its own cluster, whereas k-means starts with all terms in the same cluster and recursively splits each (sub-)cluster.

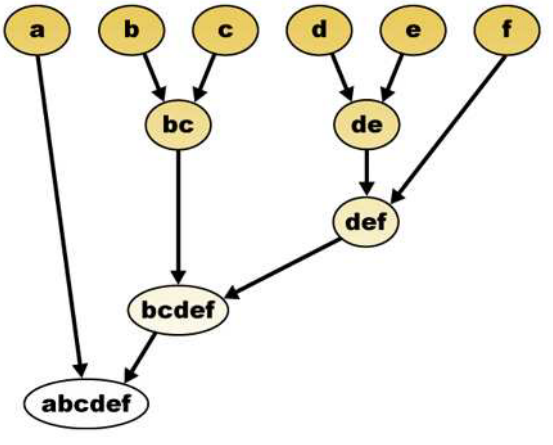

Figure 1. Structure produced by hierarchical clustering methods. Picture taken from Wikimedia Commons (http://commons.wikimedia.org), file name "Hierarchical_clustering_diagram.png".

3.1 Bottom-up agglomerative clustering

We start by building a word-space model, using the settings that gave the best results in the pre-study. We use:

- Window size: 500 (in each direction).
- Distance weighting: flat.

- Feature weighting: none.
- Left/right distinction: not made.
- Minimum feature frequency: 51.
- Dimensionality reduction: svd, 200 dimensions.

We employ a version of average linking, where we start by calculating a *centroid* representation for each cluster and then calculate the average similarity of each cluster member to this centroid.

3.2 K-means clustering

The agglomerative clustering approach, described in the previous section, produces a binary tree. Since we have many terms to cluster (1,164 terms in total – we only cluster terms with a minimum frequency of 50 in the English corpus), this results in a very deep tree, especially if we compare it with the FMA ontology, which is much flatter. Further, a binary tree will never be able to correctly capture some hierarchical relations. E.g., the relations between *finger* and *thumb*, *index finger*, *middle finger*, *ring finger* and *little finger* are not binary (one-to-one) but n-ary (one-to-many). We would like to model the relationship with *finger* directly dominating the others. Using hierarchical k-means clustering, we are no longer forced to produce binary trees; we can simply tell the algorithm how many times we would like to split the cluster at each iteration. Though we still do not get a structure where one term directly dominates other terms, but rather a one-to-many variant of the structure shown in figure 1, we at least have a chance of producing a model which is *closer* in structure to the FMA ontology.

For each clustering step, we try to find the appropriate k for splitting that particular cluster. We iterate through different values of k and evaluate each clustering by calculating the harmonic mean of *intra similarity* and *inter distance* between the clusters [15] and choose the best performing k in each step. In our experiments, we set an upper limit for k to 20, since it would be very time consuming to evaluate *every* possible k–value.[7]

3.3 Clustering from multilingual evidence

To test the effects of including evidence from more than one language when performing the clustering, we started by building four separate distributional models, one for each language, using the same settings as described in 2.3. Next, for each term in every non-English model, we look up in the FMA ontology if it is listed as a translation of any of the English terms. If it is, we concatenate the vector for this non-English term to the vector of the English term, resulting in a vector that is twice the length of the original vector. This process is repeated for every non-English language, which means that the final vectors we are working with are four times the length of the original vectors (since we are using four languages). Figure 2 illustrates the idea behind such a multilingual vector.

3.4 Evaluating the hierarchical clustering

Some of the first measures for evaluating the similarity between two ontologies (also applicable to taxonomies in general) were introduced in [11]. Further additions and alterations have been suggested since then, see [2]. In [3], an attempt is made to establish a standard measure called TF_{CSC}, which is the harmonic mean between

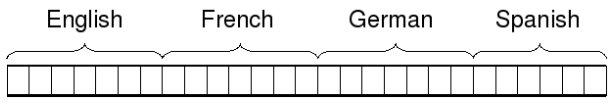

Figure 2. Distributional information from each language is concatenated to form an elongated version of the co-occurrence vector. The vectors used in the monolingual experiments consist only of the part marked 'English'.

TP_{CSC} (TP for taxonomic precision) and TR_{CSC} (TR for taxonomic recall). "CSC" stands for "common semantic cotopy", where "semantic cotopy" refers to the set of all super- and sub concepts of a particular concept and "common" means that one only takes the concepts shared by both ontologies into consideration. If O_C is the computed ontology and O_R is the reference ontology, then TP_{csc} and TR_{csc} are calculated as:

$$TP_{csc}(O_C, O_R) := \frac{1}{|O_C \cap O_R|} \sum_{c \in O_C \cap O_R} tp_{csc}(c, O_C, O_R)$$

$$TR_{csc}(O_C, O_R) := TP_{csc}(O_R, O_C)$$

$$tp_{csc}(c, O_C, O_R) := \frac{|csc(c, O_C, O_R) \cap csc(c, O_R, O_C)|}{|csc(c, O_C, O_R)|}$$

$$csc(c, O_C, O_R) := \{c_i | c_i \in O_C \cap O_R \wedge (c_i <_C c \vee c <_C c_i)\}$$

These measures have also been implemented and made available as open source and as a web service through the University of Sheffield.[8] There is a serious problem with using this evaluation in our case, however. Our gold standard, the FMA ontology, is a traditional hierarchy, where one term (concept) dominates another to form a hierarchy tree. Our automatically created results look like variants of the structure shown in figure 1. If we consider the situation more closely, we realize that the CSC of any concept between these two hierarchies will always be empty, which makes this approach unfit for evaluating our results.

To evaluate the type of taxonomies we are dealing with here, we propose to use Pearson's product-moment correlation coefficient (PMCC). The idea is that we can characterize a taxonomy by listing all pairs of concepts that it contains, along with the distance between each concept pair. E.g., for the ontology in figure 3, we get the following distances:

```
A0 -> A:    1
A0 -> A1:   2
A0 -> root: 2
A0 -> B:    3
A0 -> B0:   4
A0 -> B1:   4
A1 -> A:    1
A1 -> root: 2
A1 -> B:    3
A1 -> B0:   4
A1 -> B1:   4
A  -> root: 1
...
```

[7] Choosing 20 as upper limit as opposed to any other number was an arbitrary choice.

[8] http://wit.shef.ac.uk:8080/onteval/

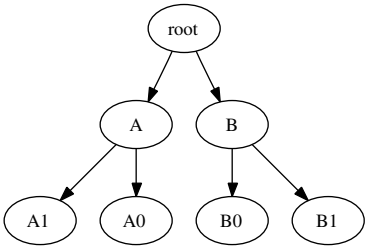

Figure 3. Small example taxonomy.

Once we have calculated these series, we can use them to calculate the PMCC measure:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where X is the series from the reference taxonomy and Y the series from the learned taxonomy, cov stands for covariance and σ is the standard deviation. The measure returns a value between -1 and 1, where 1 means perfect correlation, 0 means no correlation and -1 means perfect negative correlation, somewhat simplified.

4 Results and discussion

First off, we compare the bottom-up agglomerative clustering to the hierarchical k-means clustering described in section 3.2. Table 1 shows that the k-means approach gives a result which is substantially closer to the gold standard than the bottom-up agglomerative approach does. Since the k-means clustering uses a random initialization, we repeated the experiments ten times and report the average correlation and the standard deviation for this approach. These experiments were carried out using only the English terms and texts.

	ρ	σ	ρ_{min}	ρ_{max}
Bottom-up aggl.	0.109	N/A	0.109	0.109
K-means hierarch.	0.166	0.043	0.114	0.237

Table 1. Comparing bottom-up and k-means monolingual clustering. ρ is the correlation, σ the standard deviation.

We see two possible explanations for this improvement. One is that, since we are evaluating different k for each new split and sticking with the best one, it is possible that we this way are able to find a more "data-cohesive" way of splitting the terms. Another explanation could be that we are mimicking the flatter structure of the FMA ontology better this way, than we are with the bottom-up approach. As ever, a combination of these two factors seems most likely.

Turning our attention now to the comparison of the mono- and multilingual cases, table 2 shows that the multilingual clustering on average gives a considerable increase in correlation, paired with a marked decrease in standard deviation, indicating that this method is less sensitive to different random initializations.

Now, one might argue that these improvements are not surprising – more data is always more data. However, as was stated in the introduction, since the additional data comes from different languages than the original data, it was not self evident that the added data would help to clarify the taxonomy extraction, rather than confuse the models. Our results, however, *do* support using the multilingual evidence for this application.

	ρ	σ	ρ_{min}	ρ_{max}
Monolingual clustering	0.166	0.043	0.114	0.237
Multilingual clustering	0.201	0.027	0.137	0.229

Table 2. Comparing mono- and multilingual k-means clustering.

Previous research (see references in Sect. 1) has demonstrated the ability of distributional similarity models to capture relevant information for the task at hand and that the resulting hierarchical clustering methods do capture useful semantic information. The focus of this article therefore is not to demonstrate this once more, but rather, again based on the articles previously referred to, we take this as a given and instead investigate if the distributional models can be made even more useful by including multilingual data. This is in fact what we see confirmed in our experiments.

The approach for building the multilingual model presented here assumes that we have access to a domain-specific bilingual (or multilingual) dictionary. One could imagine getting by without such a dictionary and instead using machine translation techniques to identify term equivalents [8]. Because we are dealing with comparable rather than parallel texts here, we would have had to resort to techniques like the ones suggested by e.g., [14, 4, 5]. These have the disadvantage of being much less accurate than techniques developed for parallel texts. To avoid evaluating the quality of a term translation system rather than the effects of multilingual evidence, we decided on using the translation information coded in the FMA ontology as our lexicon. This seems not too far fetched a scenario: having access to a domain-specific bilingual dictionary and wishing to extract a taxonomy for the terms listed there.

5 Conclusions

In our experiments, we have focused on clustering based on distributional similarity. Other researchers have experimented with including other types of information for taxonomy extraction, such as two terms (NPs) sharing the same head noun [2, 16], two terms appearing in certain lexico-syntactic patterns [7] or combining the hyponymy (is-a) and cohyponymy (sibling) relations [17]. We have ongoing experiments where we include this type of information in the taxonomy extraction process and we are optimistic that the multilingual approach presented here will prove equally beneficial in these cases.

Summing up, the increase in average correlation and decrease in standard deviation when evaluating against the gold standard mean that we can make a strong case for the usefulness of multilingual evidence for the taxonomy extraction task. What's more, the resulting resource has added value when compared with the monolingual approach, since we are now free to switch between languages at will, while staying within the same taxonomic structure.

ACKNOWLEDGEMENTS

This research has been supported in part by the THESEUS Program in the MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. The responsibility for this publication lies with the authors. We also thank three anonymous reviewers for their helpful comments.

REFERENCES

[1] Stephan Bloehdorn, Philipp Cimiano, and Andreas Hotho, 'Learning ontologies to improve text clustering and classification', in *From Data and Information Analysis to Knowledge Engineering: Proceedings of the 29th Annual Conference of the German Classification Society (GfKl 2005)*, eds., Myra Spiliopoulou, Rudolf Kruse, Andreas Nürnberger, Christian Borgelt, and Wolfgang Gaul, volume 30 of *Studies in Classification, Data Analysis, and Knowledge Organization*, pp. 334–341, Magdeburg, Germany, (2006). Springer-Verlag.

[2] Philipp Cimiano, *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*, Springer-Verlag, New York, NY, USA, 2006.

[3] Klaas Dellschaft and Steffen Staab, 'On how to perform a gold standard based evaluation of ontology learning', in *5th International Semantic Web Conference*, Athens, GA, USA, (2006).

[4] Pascale Fung and Kathleen McKeown, 'Finding terminology translations from non-parallel corpora', in *The 5th Annual Workshop on Very Large Corpora*, pp. 192–202, Hong Kong, (1997).

[5] Eric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Hervé Déjean, 'A geometric view on bilingual lexicon extraction from comparable corpora', in *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pp. 526–533, Barcelona, Spain, (July 2004).

[6] Gene H. Golub and Charles F. van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, USA, 3 edn., 1996.

[7] Marti Hearst, 'Automatic acquisition of hyponyms from large text corpora', in *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France, (1992).

[8] Hans Hjelm, 'Identifying cross language term equivalents using statistical machine translation and distributional association measures', in *Proceedings of Nodalida 2007, the 16th Nordic Conference of Computational Linguistics*, Tartu, Estonia, (2007).

[9] Hans Hjelm and Christoph Schwarz, 'LiSa - morphological analysis for information retrieval', in *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, ed., Stefan Werner, volume 1 of *University of Joensuu electronic publications in linguistics and language technology*. NoDaLiDa, Ling@JoY, (2006).

[10] Christian Jacquemin, *Spotting and Discovering Terms through Natural Language Processing*, The MIT Press, Cambridge, Massachusetts, USA, 2001.

[11] Alexander Maedche, *Ontology Learning for the Semantic Web*, Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[12] Alexander Maedche, Viktor Pekar, and Steffen Staab, 'Ontology learning part one – on discovering taxonomic relations from the web', in *Web Intelligence*, eds., Ning Zhong, Jiming Liu, and Yiyu Yao, chapter 14, Springer Verlag, New York, NY, USA, (2003).

[13] Inderjeet Mani, Ken Samuel, Kris Concepcion, and David Vogel, 'Automatically inducing ontologies from corpora', in *Proceedings of CompuTerm 2004: 3rd International Workshop on Computational Terminology*, Geneva, Switzerland, (2004). COLING.

[14] Reinhard Rapp, 'Automatic identification of word translations from unrelated english and german corpora', in *Proceedings of the 37th Annual Meeting of the ACL (ACL'99)*, College Park, MD, USA, (1999).

[15] Magnus Rosell, *Clustering in Swedish – The Impact of some Properties of the Swedish Language on Document Clustering and an Evaluation Method*, Licentiate thesis, School of Computer Science and Communication, Royal Institute of Technology, Stockholm, Sweden, 2005.

[16] Pum-Mo Ryu and Key-Sun Cho, 'Taxonomy learning using term specificity and similarity', in *Proceedings from the Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge (with Coling.ACL 2006)*, pp. 41 – 48, Sydney, Australia, (2006).

[17] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng, 'Semantic taxonomy induction from heterogeneous evidence', in *Proceedings of COLING/ACL 2006*, Sydney, Australia, (2006).