Term Clustering using a Corpus-Based Similarity Measure*

Goran Nenadic, Irena Spasic, Sophia Ananiadou
Computer Science, University of Salford, UK

{G.Nenadic, I.Spasic, S.Ananiadou}@salford.ac.uk

Abstract: In this paper we present a method for the automatic term clustering. The method uses a hybrid similarity measure to cluster terms automatically extracted from a corpus by applying the C/NC value method. The measure comprises contextual, functional and lexical similarity, and it is used to instantiate the cell values in a similarity matrix. The clustering algorithm uses either the nearest neighbour or the Ward's method to calculate the distance between clusters. The approach has been tested and evaluated in the domain of molecular biology and the results are presented.

1. Introduction

The identification of concepts, linguistically represented by domain specific terms [2], is a basic step in the automated acquisition of knowledge from textual documents. Textual documents describing new knowledge in an intensively expanding domain, such as molecular biology, are swamped by new terms representing newly identified/created concepts. This makes the automatic term extraction tools essential assets for efficient knowledge acquisition. However, automatic term extraction itself is not sufficient when it comes to structuring newly acquired knowledge. Namely, the extracted terms need to be associated with other extracted terms and the terms already stored in the existing knowledge-bases. The process of linking semantically similar terms together, called term clustering, irrefutably has a positive impact on improving information extraction, information retrieval, knowledge acquisition, and document categorisation.

In this paper, we present term clustering based on the automatic discovery of term similarities [6]. The similarity measure is corpus-dependent as we base similarities on the automatic extraction of lexical and syntactical patterns in which terms appear. This measure is fed into a clustering algorithm to link similar terms.

The paper is organised as follows. Section 2 gives an overview of the term similarity measure. The clustering approach is presented in Section 3, and the results of the experiments are presented in Section 4.

2. Term Similarity Measure

We introduce a hybrid similarity measure that combines three types of term similarity measures: contextual, functional and lexical. In this section, we provide a brief overview of the three similarity measures.

Our approach to *contextual similarity* is based on automatic pattern mining, which involves the identification of the most relevant lexico-syntactic patterns that describe contexts around the terms. *Context pattern* (CP) is a lexicalised regular

* This research is part of the BioPATH research project coordinated by LION BioScience (http://www.lionbioscience.com) and funded by German Ministry of Research.

expression that corresponds to left/right context of a term. Its basic constituents are the syntactical categories of the words that constitute a term context. However, other grammatical and lexical information (e.g. the lemmatised form of a simple/compound word) can also be used to instantiate the CP constituents. Some of the CP constituents may be discarded if deemed impertinent to discriminate terms. In our experiments, we instantiated terms and verbs and removed adverbs and linking words from the CPs.

The relevance of an individual CP is determined according to a measure called CP-value. *CP-value* ranks CPs conforming to the following criteria: the frequency with which a CP occurs in a given corpus ($f(p)$), its length as the number of constituents ($|p|$), and the frequency with which it occurs as a part of other CPs ($|T_p|$, where T_p is a set of all CPs that contain p):

$$CP(p) = \begin{cases} \log_2|p| \cdot f(p); & p \text{ is not nested} \\ \log_2|p| \cdot \left(f(p) - \frac{1}{|T_p|} \sum_{b \in Tp} f(b) \right) & \text{; otherwise} \end{cases}$$

The higher the CP-value of a CP, the more relevant the CP is. Note that the relevant CPs are automatically identified. However, they are domain-specific as they rely solely on the information found in a domain specific corpus.

Contextual similarity between terms is measured by comparing the sets of CPs associated with them. Namely, if C_1 and C_2 are two sets of CPs associated with terms t_1 and t_2 respectively, then the contextual similarity between t_1 and t_2 is defined as follows:

$$CS(t_1, t_2) = \frac{2|C_1 \cap C_2|}{2|C_1 \cap C_2| + |C_1 \setminus C_2| + |C_1 \setminus C_2|}$$

In order to measure *functional similarity* between terms, we used several lexical patterns that indicate a high degree of correlation between terms. In each of these patterns terms are used concurrently within the same context. We base our approach on a hypothesis that the concurrent usage of terms within the same context indicates that the terms involved are highly correlated. Some of these patterns have been previously used to discover hyponym relations between terms [3], and some describe coordination of terms. Functional similarity between two terms equals 1, if the two terms appear concurrently in any one of the predefined lexical patterns, and 0 otherwise.

Lexical similarity between terms is based on the lexical similarity between the words of which the terms consists. If two terms share the same head, it is likely that they share the same concept as an (in)direct hypernym (e.g. `progesterone receptor` and `estrogen receptor`), and, therefore, can be regarded as being similar. Furthermore, if one of such terms has additional modifiers, then this may indicate concept specialisation (e.g. `nuclear receptor` and `orphan nuclear receptor`), and again we use this fact to treat such terms as similar. Bearing this in mind, we base the definition of lexical similarity on sharing a head and/or modifier(s). Formally, if t_1 and t_2 are terms, H_1 and H_2 their heads, and M_1 and M_2

the sets of the stems of their modifiers, then the lexical similarity between t_1 and t_2 is calculated as follows:

$$LS(t_1, t_2) = \frac{1}{a+b}\left(a * |H_1 \cap H_2| + b * \frac{2|M_1 \cap M_2|}{2|M_1 \cap M_2| + |M_1 \setminus M_2| + |M_1 \setminus M_2|} \right)$$

where a and b are weights such that $a > b$, since we give higher priority to shared heads over shared modifiers.

Finally, the hybrid term similarity measure is defined as a linear combination of the three similarity measures described above:

$$S(t_1, t_2) = \alpha\, CS(t_1, t_2) + \beta\, FS(t_1, t_2) + \gamma\, LS(t_1, t_2) \quad (1)$$

The choice of the weights α, β and γ in formula (1) is not a trivial problem. Therefore, we applied a genetic algorithm approach in order to learn the weights automatically [6]. The resulting weights were $\alpha = 0.13$, $\beta = 0.06$, and $\gamma = 0.81$. The experiments described in Section 4 are based on these values.

3. Term Clustering

Term similarities, based on the hybrid measure, are used as a basis for establishing coherent term clusters, which link semantically similar terms together. Term similarities are fed into a similarity matrix. Each row in the matrix represents a similarity vector corresponding to a specific term. The distances between such vectors are used to establish clusters. We have used hierarchical clustering based on two different clustering methods: the nearest neighbour (NN) and the Ward's method.

The distance between two clusters in the NN method [1] is determined as the minimal distance between the members of the two respective clusters. The algorithm starts with a set of clusters each containing a single term. In each step, two clusters with the minimal distance are merged. On the other hand, the Ward's method [1] aims at minimising the increase in the sum of the distances between the members of a potential cluster. In other words, the method minimises the variance within a cluster. These two methods are opposed to each other in the sense that the NN method (also known as the single linkage method) tends to produce long chain-like clusters, since the clusters are "chained" via their nearest members, while the Ward's method favours spherical clusters. In both cases, the resulting hierarchy (dendrogram) is subsequently decomposed into a set of clusters by cutting off the hierarchy at the certain depth and collecting the leaves corresponding to a sub-tree being cut off (see Figure 1).

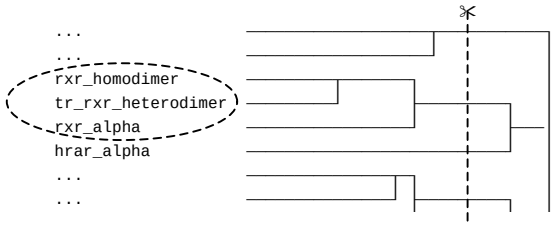

Figure 1: Producing clusters by cutting off the subtrees of the dendrogram

4. Experiments and Evaluation

Clustering techniques have been incorporated into the ATRACT workbench [5] and tested in the domain of molecular biology. The testing corpus contained 2008 abstracts retrieved from the MEDLINE database [4]. Clustering has been applied to a set of 174 top-ranked terms automatically extracted from the corpus using the C/NC method [2]. The resulting clusters have been evaluated by a domain expert, and the results, after discarding the singleton clusters, are given in Table 1. Although the distribution of clusters differed significantly for the two clustering methods, the overall precision did not significantly vary. However, the higher number of small clusters produced by the Ward's method is preferred, as the clusters are more coherent.

Cardinality of a cluster	**Nearest neighbour**			**Ward's method**		
	# of clusters	# of correct		# of clusters	# of correct	
		clusters	terms		clusters	terms
2	16	7 (44%)	14	33	22 (67%)	44
3	7	6 (86%)	18	19	10 (53%)	30
4	4	2 (50%)	8	5	3 (60%)	12
≥ 5	10	7 (70%)	47	2	1 (50%)	8
Total:	**37**	**22 (59%)**	**87 (63%)**	**59**	**36 (61%)**	**114 (71%)**

Table 1: Clustering results

5. Conclusion

We have presented the results on term clustering using a hybrid term similarity measure. The measure is based on lexical and syntactical patterns automatically extracted from a corpus. The method achieves around 70% precision in clustering semantically similar terms. It also proved to be consistent as similar terms shared most of their "friends". Since the initial results are promising, we plan to improve the results by further investigation into the clustering methods, the hybrid similarity measure and the size of corpus, since the measure is corpus-dependant.

References

1. Fasulo D. (1999) *Analysis on Recent Work on Clustering Algorithms*. Technical Report #01-03-02, Dept. of Computer Science and Engineering, University of Washington, Seattle, p. 24.
2. Frantzi K.T., Ananiadou S. and Mima H. (2000) *Automatic Recognition of Multi-Word Terms: The C-value/NC-value Method.* International Journal on Digital Libraries, 3/2, pp. 115-130.
3. Hearst M.A. (1992) *Automatic Acquisition of Hyponyms From Large Text Corpora*. Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France.
4. MEDLINE (2002) *National Library of Medicine*. Available at: http://www.ncbi.nlm.nih.gov/PubMed/
5. Mima H., Ananiadou S. and Nenadic G. (2001) *ATRACT Workbench: An Automatic Term Recognition and Clustering of Terms.* Text, Speech and Dialogue - TSD 2001, LNAI 2166, Springer-Verlag, Berlin, pp. 126-133.
6. Spasic I., Nenadic G., Manios K. and Ananiadou S. (2002) *Supervised Learning of Term Similarities.* Submitted to IDEAL '02, Manchester, UK.