Phrase clustering without document context.

Eric Sanjuan, Fidelia Ibekwe-Sanjuan

Phrase Clustering without document context

Eric SanJuan[1] and Fidelia Ibekwe-SanJuan[2]

[1] LITA, University of Metz – URI, INIST-CNRS, France
eric.sanjuan@univ-metz.fr
[2] URSIDOC – University of Lyon 3, France
fidelia.ibekwe@univ-lyon3.fr

Abstract. We applied different clustering algorithms to the task of clustering multi-word terms in order to reflect a humanly built ontology. Clustering was done without the usual document co-occurrence information. Our clustering algorithm, CPCL (Classification by Preferential Clustered Link) is based on general lexico-syntactic relations which do not require prior domain knowledge or the existence of a training set. Results show that CPCL performs well in terms of cluster homogeneity and shows good adaptability for handling large and sparse matrices.

1 Introduction

We test the ability of clustering methods in an *out-of-context clustering* (OTC) task, i.e., clustering without document co-occurrence information. The methods are evaluated against categories issuing from a humanly built ontology. For this purpose, we chose as test corpus the GENIA dataset which comes with an existing *ideal partition*. Domain terms in this corpus have been manually annotated by specialists, yielding $31,398$ terms. The GENIA ontology consists of 36 categories at the leaf nodes. Each term in the GENIA corpus has been assigned a semantic category at the leaf node of the ontology. The goal of the evaluation is to determine the method whose output requires the least effort to reproduce the categories at the leaf nodes of the ontology.

2 Our clustering methodology

We developed a fast and efficient clustering algorithm, CPCL that builds clusters of multi-word terms (MWTs) without relying on document context. Details of our clustering methodology can be found in [1]. Here we only sketch out its principle. Terms are clustered depending on the presence and number of shared lexico-syntatic relations. Two types of lexico-syntatic operations are studied: the expansion of an existing term by the addition of one or more modifier words (*information retrieval – efficient retrieval of information*); the substitution a word in a term, either in the modifier position (*coronary heart disease – coronary lung disease*) or in the head position (*mutant motif – mutant strain*). We call $COMP$ the subset of relations that affects modifier words in a term and $CLAS$ the subset that affects the head word in a term. Clustering is based on $COMP$ and $CLAS$ relations and CPCL, a graph-based algorithm called which implements a variant of hierarchical clustering. Let us refer to this principle of

clustering as "clustering by lexico-semantic similarity" (LSS). $COMP$ relations are used in an initial phase to form connected components and $CLAS$ relations are used in the 2nd phase to form clusters of such components in a hierarchical process. The particularity of $CPCL$ is to compute at each iteration the local maximal similarity values in the graph of non null similarity relations. Average link clustering is then performed on the resulting subgraph.

3 Evaluation metrics

For the OTC task, we need a measure that focuses on cluster quality (homogeneity) vis-à-vis an existing partition (here the GENIA categories) and that is also adapted to the comparison of methods producing a great number of clusters (hundreds or thousands) and of very differing sizes. Pantel & Lin's editing distance [2] appears as the most suitable for this task. We focus on two of the elementary operations in their measure: "merges" which is the union of disjoint sets and "moves" that applies to singular elements. In this restricted context, Pantel & Lin's measure has a more deterministic behaviour with some inherent bias which we correct hereafter.

Let Ω be a set of objects for which we know a crisp classification $\mathcal{C} \subseteq 2^{\Omega}$. Consider now a second disjoint family $\mathcal{F}$ of subsets of Ω representing the output of a clustering algorithm. For each cluster $F \in \mathcal{F}$, we denote by $\mathcal{C}_F$ the class $C \in \mathcal{C}$ such that $|C \cap F|$ is maximal. We thus propose a corrected version of this measure where the weight of each move is no more 1 but $|\Omega|/(|\Omega| - \max\{|C| : C \in \mathcal{C}\})$ and the weight of a merge is $|\Omega|/(|\Omega| - |\mathcal{C}|)$:

$$\mu_{ED}(\mathcal{C}, \mathcal{F}) = 1 - \frac{\max\{0, |\mathcal{F}| - |\mathcal{C}|\}}{|\Omega| - |\mathcal{C}|} - \frac{\sum_{F \in \mathcal{F}}(|F| - |\mathcal{C}_F \cap F|)}{|\Omega| - \max\{|C| : C \in \mathcal{C}\}} \quad (1)$$

The maximal value of μ_{ED} is 1 in the case where the clustering output corresponds exactly to the target partition. It is equal to 0 in the case that $\mathcal{F}$ is a trivial partition (discrete or complete). Based on the corrected μ_{ED} index, we propose a complementary index, *cluster homogeneity* (μ_H) defined as:

$$\mu_H(\mathcal{C}, \mathcal{F}) = \frac{\mu_{ED}(\mathcal{C}, \mathcal{F})}{1 + \sum_{F \in \mathcal{F}}(|F| - |\mathcal{C}_F \cap F|)} \times |\Omega| \quad (2)$$

μ_H takes its maximal value $|\Omega|$ if $\mathcal{F} = \mathcal{C}$ and, like the μ_{ED} measure, it is null if $\mathcal{F}$ is one of the two trivial partitions. We will use μ_H to distinguish between algorithms having similar editing distances but not producing clusters of the same quality.

4 Experimental setup

For statistical clustering methods to find sufficient *co-occurrence* information, it was necessary to represent *term-term* similarity. Co-occurrence is defined here as internal word co-occurrence within a term. We then built a *term* × *word* matrix where the rows were the terms and the columns the unique constituent words. We further adapted this matrix as follows: words are assigned a weight according to their grammatical role in the term and their position with regard to the head word. Since a head word is the noun focus (the subject), it receives

a weight of 1. Modifier words are assigned a weight which is the inverse of their position with regard to the head word. Let M be the term×word matrix such that $M_{i,j}$ refers to the weight of word j in term i. We derive two other matrices from M. A similarity matrix $S = M.M^t$ whose cells give the similarity between two terms as the scalar product of their vectors (for hierarchical algorithms). A core matrix C for partitioning algorithms by removing all rows and columns of M with less than 5% of non null values.

We experimented three types of clustering relations on four clustering methods. The three clustering relations were:

- Coarse Lexical Similarity (CLS). This consists in grouping terms by identical head word and will serve as a "baseline" against which the other algorithms can be aligned.
- Lexico-Syntactic Similarity (LSS). This is based on the linguistic relations identified by our clustering methodology as described in section §2.
- Lexical Cohesion (LC). This is based on the vector representation of terms in the space of words they contain as described in section §4.

The following clustering algorithms were tested:

- **Baseline with CLS**: No particular parameter is necessary. All terms sharing the same head word are grouped in the same cluster.
- **CPCL with LSS**: Custering is based on LSS relations. No threshold was set so as not to exclude terms and relations. The algorithm was stopped at iteration 1. We also tested the performance of the 1st step of CPCL, i.e., the connected components formed at the $COMP$ level.
- **Hierarchical with LC:** Clustering is based on the similarity matrix $S[S \geq th]$ where th is a threshold with the following values: 0.5 and 0.8.
- **Partitioning with LC:** This is based on the computation of k-means centers and medoids on the core matrix C. We used the standard functions of k-means and CLARA (Clustering LARge Applications).We ran these two variants for the following values of k: 36, 100, 300, 600 and 900.

5 Results

The baseline clustering grouped the whole list of terms in $3,220$ clusters. CPCL on LSS generated $1,897$ non trivial components at the COMP phase and $3,738$ clusters at the CLAS phase. Hierarchical clustering on LC, based on similarity matrix generated $1,090$ clusters for a threshold of $th = 0.5$ and $1,217$ clusters for $th = 0.8$.

The hierarchical algorithm with *th=0.8* and CPCL obtain a better μ_{ED} score (≥ 0.36) than the baseline (≤ 0.24) and partitioning methods (≤ 0.14) when considering all terms (length ≤ 2). When fewer and longer terms are considered (length ≥ 3), partitioning methods obtain μ_{ED} scores between 0.58 and 0.79 and outperform the baseline, CPCL and hierarchical algorithms (≤ 0.59). However, the μ_{ED} measure masks important features of the evaluation: how homogeneous a cluster is with regard a category in the target partition.

Cluster homogeneity is measured by the μ_H index which computes the ratio between the value of μ_{ED} and the number of movings. This is plotted on figure 1.

Since the majority of the clustering methods showed sensitivity to term length, we plotted the score obtained by each of the measure (y-axis) by term length (x-axis). Note that at each length, only terms of that length and above are considered. Thus, the further we move down the x-axis, the fewer the input terms for clustering. The baseline clustering is noted "basic" on this figure.

It appears clearly that CPCL outperforms the other methods. It forms the most homogeneous clusters that need the least number of moves and merges in order to obtain the target partition. Also, CPCL is the only algorithm that significantly outperforms the baseline, irrespective of term length.

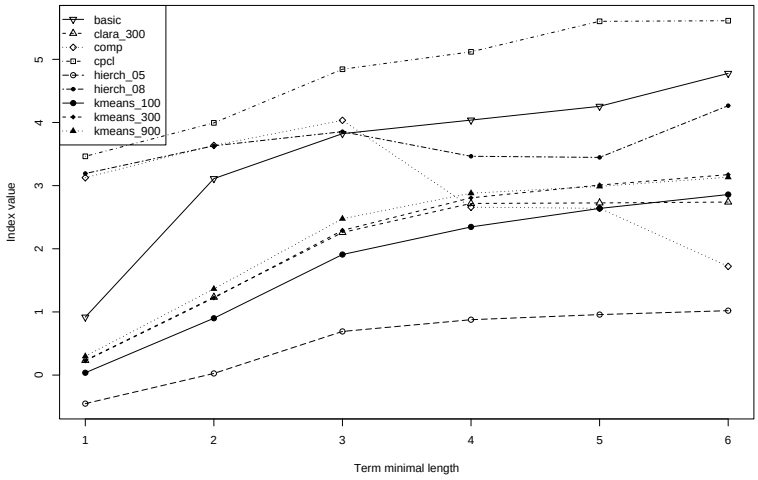

Fig. 1. Cluster homogeneity measure μ_{ED}.

6 Conclusion

Overall, this experiment has shown that even without adequate context (document co-occurrence), clustering algorithms can be adapted to partially reflect a human semantic organisation of scientific concepts. Moreover, clustering based on simple linguistic relations outperforms other criteria in terms of cluster quality.

References

1. SanJuan, E., Dowdall, J., Ibekwe-SanJuan, F., Rinaldi, F.: A symbolic approach to automatic multiword term structuring. Computer Speech and Language **19**(4) (2005) 524 – 542
2. Pantel, P., Lin, D.: Clustering by Committee. In: Annual International conference of ACM on Research and Development in Information retrieval - ACM SIGIR, Tampere, Finland (2002) 199–206