Cooperating Techniques for Extracting Conceptual Taxonomies from Text

S. Ferilli[1,2], F. Leuzzi[1], and F. Rotella[1]

[1] Dipartimento di Informatica – Università di Bari
ferilli@di.uniba.it {fabio.leuzzi, rotella.fulvio}@gmail.com
[2] Centro Interdipartimentale per la Logica e sue Applicazioni – Università di Bari

Abstract. The current abundance of electronic documents requires automatic techniques that support the users in understanding their content and extracting useful information. To this aim, it is important to have conceptual taxonomies that express common sense and implicit relationships among concepts. This work proposes a mix of several techniques that are brought to cooperation for learning them automatically. Although the work is at a preliminary stage, interesting initial results suggest to go on extending and improving the approach.

1 Introduction

The spread of electronic documents and document repositories has generated the need for automatic techniques to understand and handle the documents content, in order to help the user in satisfying his information needs without being overwhelmed by the huge amount of available data. Since most of these data are in textual form, and since text explicitly refers to concepts, most work has focussed on Natural Language Processing (NLP). Obtaining automatically *Full Text Understanding* is not trivial, due to the intrinsic ambiguity of natural language and to the huge amount of required common sense and linguistic/conceptual background knowledge. Nevertheless, even small portions of such a knowledge may significantly improve understanding performance, at least in limited domains. Although standard tools, techniques and representation formalisms are still missing, lexical and/or conceptual taxonomies can provide a useful support to many NLP tasks. Unfortunately, manually building this kind of resources is very costly and error-prone, which is a strong motivation towards automatic construction of conceptual networks by mining large amounts of documents in natural language. This work aims at partially simulating some human abilities in this field, such as extracting the concepts expressed in given texts and assessing their relevance; obtaining a practical description of the concepts underlying the terms, which in turn would allow to generalize concepts having similar descriptions; and applying some kind of reasoning 'by association', that looks for possible indirect connections between two identified concepts. The system will take in input texts in natural language, and process them to build a conceptual network that supports the above objectives. The resulting network can be visualized by means of a suitable interface and translated into a First-Order Logic

(FOL) formalism, to allow the subsequent exploitation of logic inference engines in applications that use that knowledge.

Our proposal consists in a mix of existing tools and techniques, that are brought to cooperation in order to reach the above objectives, extended and supported by novel techniques when needed. The next section briefly recalls related work. Then, Section 3 describes the mixed approach, and discusses the novel parts in more detail. A preliminary evaluation of the proposal is reported in Section 4, while Section 5 concludes the paper and outlines future work issues.

2 Related Work

Many works exist aimed at building taxonomies and ontologies from text. A few examples: [10, 9] build ontologies from natural language text by labelling the taxonomical relations only, while we label also non-taxonomical ones with actions (verbs); [14] builds a taxonomy considering only concepts that are present in a domain but do not appear in others, while we are interested in all recognized concepts independently of their being generic or domain-specific; [13] defines a language to build formal ontologies, while we are interested in the lexical level.

As regards our proposal, a first functionality that we needed is syntactic analysis of the input text. We exploited the *Stanford Parser* and *Stanford Dependencies* [7, 1], two very effective tools that can identify the most likely syntactic structure of sentences (including active and passive forms), and label their components as 'subject' or '(direct/indirect) object'. Moreover, they normalize the words in the input text using lemmatization instead of stemming, which allows to distinguish their grammatical role and is more comfortable to read by humans. We also exploited the Weka project [5], that provides a set of tools to carry out several learning and Data Mining (DM) tasks, including clustering, classification, regression, discovery of association rules and visualization.

Another technique that inspired our work is the one described in [8] to semi-automatically extract a domain-specific ontology from free text, without using external resources but focussing the analysis on *Hub Words* (i.e., words having high frequency). After building the ontology, the adaptation of a Hub Word t is ranked according to its 'Hub Weight':

$$W(t) = \alpha w_0 + \beta n + \gamma \sum_{i=1}^{n} w(t_i)$$

where w_0 is a given initial weight, n is the number of relationships in which t is involved, $w(t_i)$ is the $tf{*}idf$ weight of the i-th word related to t, and $\alpha+\beta+\gamma = 1$.

A task aimed at identifying most important words in a text, to be used as main concepts for inclusion in the taxonomy, is *Keyword Extraction* (KE). Among the several proposals available in the literature, we selected two techniques that can work on single documents (rather than requiring a whole corpus) and are based on different and complementary approaches, so that they can together provide an added value. The quantitative approach in [12] is based on the assumption that the relevance of a term in a document is proportional to how frequently it co-occurs with a subset of most frequent terms in that document. The χ^2 statistic is exploited to check whether the co-occurrences establish a significant deviation from chance. To improve orthogonality, the reference frequent terms are preliminarily grouped exploiting *similarity-based* clustering (using similar distribution of co-occurrence with other terms) and *pairwise* clustering (based on frequent co-occurrences). The qualitative approach in [3], based on WordNet [2] and its extension WordNet Domains [11], focusses on the meaning of terms instead of their frequency and determines keywords as terms associated to the concepts referring to the main subject domain discussed in the text. It exploits a density measure that determines how much a term is related to different concepts (in case of polysemy), how much a concept is associated to a given domain, and how relevant a domain is for a text.

Lastly, we need in some steps of our technique to assess the similarity among concepts in a given conceptual taxonomy. A classical, general measure, is the *Hamming distance* [6], that works on pairs of equal-lenght vectorial descriptions and counts the minimum number of changes required to turn one into the other. Other measures, specific for conceptual taxonomies, are sf_{Fa} [4] (that adopts a global approach based on the whole set of hypernyms) and sf_{WP} [16] (that focuses on a particular path between the nodes to be compared).

3 Proposed Approach

In the following, we will assume that each term in the text corresponds to an underlying concept (phrases can be preliminarily extracted using suitable techniques, and handled as single terms). A *concept* is described by a set of characterizing attributes and/or by the concepts that interact with it in the world described by the corpus. The outcome is a graph, where nodes are the concepts recognized in the text, and edges represent the relationships among these nodes, expressed by verbs in the text (whose direction denotes their role in the relationship). This can be interpreted as a *semantic network*.

3.1 Identification of Relevant Concepts

The input text is preliminarily processed by the Stanford Parser in order to extract the syntactic structure of the sentences that make it up. In particular, we are interested only in (active or passive) sentences of the form *subject-verb-(direct/indirect)complement*, from which we extract the corresponding triples ⟨*subject*, *verb*, *complement*⟩ that will provide the concepts (the *subject*s and *complement*s) and attributes (*verb*s) for the taxonomy. Indirect complements are treated as direct ones, by embedding the corresponding preposition into the verb: e.g., *to put*, *to put on* and *to put across* are considered as three different verbs, and sentence *John puts on a hat* returns the triple ⟨John,put_on,hat⟩, in which *John* and *hat* are concepts associated to attribute *put_on*, indicating that *John* can *put_on* something, while a *hat* can be *put_on*). Triples/sentences involving verb 'to be' or nouns with adjectives provide immediate hints to build the subclass structure in the taxonomy: for instance, "The dog is a domestic animal..." yields the relationships is_a(dog, animal) and is_a(domestic_animal,animal).

The whole set of triples is represented in a *Concepts*×*Attributes* matrix $\mathcal{V}$ that recalls the classical *Terms*×*Documents* Vector Space Model (VSM) [15]. The matrix is filled according to the following scheme (resembling $tf \cdot idf$):

$$\mathcal{V}_{i,j} = \frac{f_{i,j}}{\sum_k f_{k,j}} \cdot \log \frac{|A|}{|\{j : c_i \in a_j\}|}$$

where:

- $f_{i,j}$ is the frequency of the i-th concept co-occurring with the j-th attribute;
- $\sum_k f_{k,j}$ is the sum of the co-occurrences of all concepts with the j-th attribute;
- A is the entire set of attributes;
- $|\{j : c_i \in a_j\}|$ is the number of attributes with which the concept c_i co-occurrs (i.e., for which $f_{i,j} \neq 0$).

Its values represent the term frequency *tf*, as an indicator of the relevance of the term in the text at hand (no *idf* is considered, to allow the incremental addition of new texts without the need of recomputing this statistic).

A clustering step (typical in Text Mining) can be performed on $\mathcal{V}$ to identify groups of elements having similar features (i.e., involved in the same verbal relationships). The underlying idea is that concepts belonging to the same cluster should share some semantics. For instance, if concepts *dog, John, bear, meal, cow* all share attributes *eat, sleep, drink, run*, they might be sufficiently close to each other to fall in the same cluster, indicating a possible underlying semantic (indeed, they are all animals). Since the number of clusters to be found is not known in advance, we exploit the EM clustering approach provided by Weka based on the Euclidean distance applied row vectors representing concepts in $\mathcal{V}$.

Then, the application on the input texts of various Keyword Extraction techniques, based on different (and complementary) aspects, perspectives and theoretical principles, allows to identify relevant concepts. We use the quantitative approach based on co-occurrences k_c [12], the qualitative one based on WordNet k_w [3] and a psychological one based on word positions k_p. The psychological approach is novel, and is based on the consideration that humans tend to place relevant terms/concepts toward the start and end of sentences and discourses, where the attention of the reader/listener is higher. In our approach, the chance of a term being a keyword is assigned simply according to its position in the sentence/discourse, according to a mixture model determined by mixing two Gaussian curves whose peaks are placed around the extremes of the portion of text to be examined.

The information about concepts and attributes is exploited to compute a *Relevance Weight* $W(\cdot)$ for each node in the network. Then, nodes are ranked by decreasing *Relevance Weight*, and a suitable cutpoint in the ranking is determined to distinguish relevant concepts from irrelevant ones. We cut the list at the first item c_k in the ranking such that:

$$W(c_k) - W(c_{k+1}) \geq p \cdot \max_{i=0,\ldots,n-1} (W(c_i) - W(c_{i+1}))$$

i.e., the difference in relevance weight from the next item is greater or equal than the maximum difference between all pairs of adjacent items, smoothed by a user-defined parameter $p \in [0, 1]$.

Computation of Relevance Weight Identifying key concepts in a text is more complex than just identifying keywords. Inspired to the Hub Words approach, we compute for each extracted concept a *Relevance Weight* expressing its importance in the extracted network, by combining different values associated to different perspectives: given a node/concept $\overline{c}$,

$$W(\overline{c}) = \alpha \frac{w(\overline{c})}{\max_c w(c)} + \beta \frac{e(\overline{c})}{\max_c e(c)} + \gamma \frac{\sum_{(c,\overline{c})} w(c)}{e(\overline{c})} + \delta \frac{d_M - d(\overline{c})}{d_M} + \epsilon \frac{k(\overline{c})}{\max_c k(c)}$$

where α, β, γ, δ, ϵ are weights summing up to 1, and:

- $w(c)$ is an initial weight assigned to node c;
- $e(c)$ is the number of edges of any kind involving node c;
- $(c, \overline{c})$ denotes an edge involving node $\overline{c}$;
- d_M is the largest distance between any two nodes in the whole vector space;
- $d(c)$ is the distance of node c from the center of the corresponding cluster;
- $k(c)$ is the keyword weight associated to node c.

The first term represents the *initial weight* provided by $\mathcal{V}$, normalized by the maximum initial weight among all nodes. The second term considers the *number of connections* (edges) of any category (verbal or taxonomic relationships) in which $\overline{c}$ is involved, normalized by the maximum number of connections of any node in the network. The third term (*Neighborhood Weight Summary*) considers the average initial weight of all neighbors of $\overline{c}$ (just summing up the weights the final value would be proportional to the number of neighbors, that is already considered in the previous term). The fourth term represents the *Closeness to Center* of the cluster, i.e. the distance of $\overline{c}$ from the center of its cluster, normalized by the maximum distance between any two instances in the whole vector space. The last term takes into account the outcome of the three KE techniques on the given text, suitably weighted:

$$k(c) = \zeta k_c(c) + \eta k_w(c) + \theta k_p(c)$$

where ζ, η and θ are weights ranging in $[0, 1]$ and summing up to 1. These terms were designed to be independent of each other. A partial interaction is present only between the second and the third ones, but is significantly smoothed due to the applied normalizations.

3.2 Generalization of Similar Concepts

To generalize two or more concepts (G generalizes A if anything that can be labeled as A can be labeled as G as well, but not *vice-versa*), we propose to exploit WordNet and use the set of connections of each concept with its direct neighbors as a description of the underlying concept. Three steps are involved in this procedure:

1. *Grouping similar concepts*, in which all concepts are grossly partitioned to obtain subsets of similar concepts;
2. *Word Sense Disambiguation*, that associates a single synset to each term by solving possible ambiguities using the domain of discourse (Algorithm 1);
3. *Computation of taxonomic similarity*, in which WordNet is exploited to confirm the validity of the groups found in step 1 (Algorithm 2).

As to step 1, we build a *Concepts*×*Concepts* matrix $\mathcal{C}$ where $\mathcal{C}_{i,j} = 1$ if there is at least a relationship between concepts i and j, or $\mathcal{C}_{i,j} = 0$ otherwise. Each row in $\mathcal{C}$ can be interpreted as a description of the associated concept in terms of its relationships to other concepts, and exploited for applying a pairwise clustering procedure based on *Hamming distance*. In detail, for each possible pair of different row and column items whose corresponding row and column are not null and whose similarity passes a given threshold: if neither is in a cluster yet, a new cluster containing those objects is created; otherwise, if either item is already in a cluster, the other is added to the same cluster; otherwise (both already belong to different clusters) their clusters are merged. Items whose similarity with all other items does not pass the threshold result in singleton clusters.

This clustering procedure alone might not be reliable, because terms that occur seldom in the corpus have few connections (which would affect their cluster assignment due to underspecification) and because the expressive power of this formalism is too low to represent complex contexts (which would affect even more important concepts). For this reason, the support of an external resource might be desirable. We consider WordNet as a sensible candidate for this, and try to map each concept in the network to the corresponding synset (a non trivial problem due to the typical polysemy of many words) using the *one domain per discourse* assumption as a simple criterion for Word Sense Disambiguation: the meanings of close words in a text tend to refer to the same domain, and such a domain is probably the dominant one among the words in that portion of text. Thus, WordNet allows to check and confirm/reject the similarity of concepts belonging to the same cluster, by considering all possible pairs of words whose similarity is above a given threshold. The pair (say $\{A, B\}$) with largest similarity value is generalized with their most specific common subsumer (hypernym) G in WordNet; then the other pairs in the same cluster that share at least one of the currently generalized terms, and whose least common hypernym is again G, are progressively added to the generalization. Similarity is determined using a mix of the measures proposed in [4] and in [16], to consider both the global similarity

Algorithm 1 Find “best synset” for a word

```
Input: word $t$, list of domains with weights.
Output: best synset for word $t$.

  $best\_synset \leftarrow empty$
  $best\_domain \leftarrow empty$
  for all $synset(s_t)$ do
    $max\_weight \leftarrow -\infty$
    $optimal\_domain \leftarrow empty$
    for all $domains(d_s)$ do
      if $weight(d_s) > max\_weight$ then
        $max\_weight \leftarrow weight(d_s)$
        $optimal\_domain \leftarrow d_s$
      end if
    end for
    if $max\_weight > weight(best\_domain)$ then
      $best\_synset \leftarrow s_t$
      $best\_domain \leftarrow optimal\_domain$
    end if
  end for
```

and the actual viability of the specific candidate generalization:

$$sf(A,B) = sf_{Fa}(A,B) \cdot sf_{WP}(A,B)$$

3.3 Reasoning ‘by association’

Reasoning ‘by association’ means finding a path of pairwise related concepts that establishes an indirect interaction between two concepts c' and c'' in the semantic network. We propose to look for such a path using a *Breadth-First Search* (BFS) technique, applied to both concepts under consideration.The expansion steps of the two processes are interleaved, checking at each step whether the new set of concepts just introduced has a non-empty intersection with the set of concepts of the other process. When this happens, all the concepts in such an intersection identify one or more shortest paths connecting c' and c'', that can be retrieved by tracing back the parent nodes at each level in both directions up to the roots c' and c''. Since this path is made up of concepts only, to obtain a more sensible ‘reasoning’ it must be filled with the specific kind of interaction represented by the labels of edges (verbs) that connect adjacent concepts in the chain.

4 Evaluation

The proposed approach was evaluated using *ad-hoc* tests that may indicate its strengths and weaknesses. Due to lack of space, only a few selected outcomes will be reported here. Although preliminary, these results seem enough to suggest that the approach is promising. The following default weights for the *Relevance Weight* components were empirically adopted:

- $\alpha = 0.1$ to increase the impact of most frequent concept (according to tf);

Algorithm 2 Effective generalization research.

Input: the set of C clusters returned by pair-wise clustering; T similarity threshold.
Output: set of candidate generalizations.

```
$generalizations \leftarrow empty\ set$
for all $c \in C$ do
  $good\_pairs \leftarrow empty\ set$
  for all $pair(O_i, O_j) \mid i, j \in c$ do
    if $similarity\_score(pair(O_i, O_j)) > T$ then
      $good\_pairs.add(pair(O_i, O_j), wordnet\_hypernym(pair(O_i, O_j)))$
    end if
    if $good\_pairs \neq empty\ set$ then
      $new\_set \leftarrow \{good\_pairs.getBestPair, good\_pairs.getSimilarPairs\}$
      $generalizations.add(new\_set)$
    end if
  end for
end for
```

good_pairs → all pairs that passed T, with the most specific common hypernym discovered in WordNet

good_pairs.getBestPair → the pair that has the best similarity score.

good_pairs.getSimilarPairs → the pairs that involve one of two objects of the best pair, that have satisfied the similarity score and have the same hypernym as the best pair

wordnet_hypernym → the most specific common hypernym discovered in WordNet for the two passed object.

- $\beta = 0.1$ to keep low the impact of co-occurrences between nodes;
- $\gamma = 0.3$ to increase the impact of less frequent nodes if they are linked to relevant nodes;
- $\delta = 0.25$ to increase the impact of the clustering outcome;
- $\epsilon = 0.25$ as for δ, to increase the impact of keywords.

while those for the KE techniques were taken as $\zeta = 0.45$, $\eta = 0.45$ and $\theta = 0.1$ (to reduce the impact of the psychological perspective, that is more naive compared to the others).

4.1 Recognition of relevant concepts

We exploited a dataset made up of documents concerning *social networks* of socio-political and economic subject. Table 1 shows on the top the settings used for three different runs, concerning the *Relevance Weight* components:

$$W = A + B + C + D + E$$

and the cutpoint value for selecting relevant concepts. The corresponding outcomes (at the bottom) show that the default set of parameter values yields 3 relevant concepts, having very close weights. Component D determines the inclusion of the very unfrequent concepts (see column A) *access* and *subset* (0.001 and 6.32 E-4, respectively) as relevant ones. They benefit from the large initial weight of *network*, to which they are connected. Using the second set of parameter values, the predominance of component A in the overall computation,

Table 1. Three parameter choices and corresponding outcome of relevant concepts.

Test #	α	β	γ	δ	ϵ	p
1	0.10	0.10	0.30	0.25	0.25	1.0
2	0.20	0.15	0.15	0.25	0.25	0.7
3	0.15	0.25	0.30	0.15	0.15	1.0

Test #	Concept #	A	B	C	D	E	W
1	network	0.100	0.100	0.021	0.178	0.250	0.649
	access	0.001	0.001	0.154	0.239	0.250	0.646
	subset	6.32E-4	0.001	0.150	0.239	0.250	0.641
2	network	0.200	0.150	0.0105	0.178	0.250	0.789
3	network	0.150	0.25	0.021	0.146	0.150	0.717
	user	0.127	0.195	0.022	0.146	0.150	0.641
	number	0.113	0.187	0.022	0.146	0.150	0.619
	individual	0.103	0.174	0.020	0.146	0.150	0.594

Table 2. Pairwise clustering statistics.

Dataset	MNC (0.001)	MNC (0.0001)	Vector size
B	3	2	1838
P	3	1	1599
$B + P$	5	1	3070

and the cutpoint threshold lowered to 70%, cause the frequency-based approach associated to the initial weight to give neat predominance to the first concept in the ranking. Using the third set of parameter values, the threshold is again 100% and the other weights are such that the frequency-based approach expressed by component A is balanced by the number of links affecting the node and by the weight of its neighbors. Thus, both nodes with highest frequency and nodes that are central in the network are considered relevant. Overall, concept *network* is always present, while the other concepts significantly vary depending on the parameter values.

4.2 Concept Generalization

Two toy experiments are reported for concept generalization. The maximum threshold for the *Hamming distance* was set to 0.001 and 0.0001, respectively, while the minimum threshold of *taxonomic similarity* was fixed at 0.4 in both. Two datasets on *Social networks* were exploited: a book (B) and a collection of scientific papers (P) concerning socio-political and economic discussions. Observing the outcome, three aspects can be emphasized: the level of detail of the concept descriptions that in pairwise clustering satisfy the criterion, the intuitivity of the generalizations supported by *WordNet Domains*, and the values of the single conceptual similarity measures applied to synsets in WordNet.

In Table 2, MNC is the *Max Number of Connections* detected among all concept descriptions that have been agglomerated at least once in the pairwise clustering. Note that all descriptions which have never been agglomerated, are

Table 3. Generalizations for different pairwise clustering thresholds (Thr.) and minimum similarity threshold 0.4 (top), and corresponding conceptual similarity scores (bottom).

Thr.	Dataset	Subsumer	Subs. Domain	Concepts	Conc. Domain
0.001	B	parent [110399491]	person	adopter [109772448] dad [109988063]	factotum person
	P	human action [100030358]	factotum	discussion [107138085] judgement [100874067]	factotum law
	$B+P$	dr. [110020890]	medicine	psychiatrist [110488016] abortionist [109757175] specialist [110632576]	medicine medicine medicine
0.0001	B	physiological state [114034177]	physiology	dependence [114062725] affliction [114213199]	physiology medicine
	P	mental attitude [106193203]	psychology	marxism [106215618] standpoint [106210363]	politics factotum
	$B+P$	feeling [100026192]	psychological features	dislike [107501545] satisfaction [107531255]	psychological features psychological features

#	Pairs	Fa score	WP score	Score
1	adopter, dad	0.733	0.857	0.628
2	discussion, judgement	0.731	0.769	0.562
3	psychiatrist, abortionist	0.739	0.889	0.657
	psychiatrist, specialist	0.728	0.889	0.647
4	dependence, affliction	0.687	0.750	0.516
5	marxism, standpoint	0.661	0.625	0.413
6	dislike, satisfaction	0.678	0.714	0.485

considered as single instance in a separated cluster. Hence, the concepts recognized as similar have very few neighbors, suggesting that concepts become ungeneralizable as their number of connections grows. Although in general this is a limitation, such a cautious behavior is to be preferred until an effective generalization technique is provided, that ensures the quality of its outcomes (wrong generalizations might spoil subsequent results in cascade).

It is worth emphasizing that not only sensible generalizations are returned, but their domain is also consistent with those of the generalized concepts. This happens with both thresholds (0.001 and 0.0001), that return respectively 23 and 30 candidate generalizations (due to space limitations, Table 3 reports only a representative sample, including a generalization for each dataset used). Analyzing the two conceptual similarity measures used for generalization reveals that, for almost all pairs, both yield very high values, leading to final scores that neatly pass the 0.4 threshold, and sf_{WP} is always greater than sf_{Fa}. Since the former is more related to a specific path, and hence to the goodness of the chosen subsumer, this confirms the previous outcomes (suggesting that the chosen subsumer is close to the generalized concepts). In the sample reported in Table 3, only case 5 disagrees with these considerations.

4.3 Reasoning by association

Table 4 shows a sample of outcomes of reasoning by association. E.g., case 5 explains the relationship between *freedom* and *internet* as follows: the adult

Table 4. Exampes of reasoning by associations (start and target nodes in emphasis).

#	Subject	Verb	Complement
1	*flexibility*	convert	life
	people	settle, desire, do_at, extinguish	life
	people	use, revolution	myspace
	myspace	develop	*headline*
2	people	member	*threat*
	people	erode, combine	technology
	computer	acknowledge	technology
3	internet	extend	*neighbor*
	majority	erode	internet
	majority	erode, do	*facebook*
4	*adult*	use	platform
	facebook	acknowledge	platform
5	adult	write	*freedom*
	adult	use	platform
	technology	acknowledge	platform
	internet	acknowledge	technology

write about freedom, and use platform, that is recognized as a technology, as well as internet.

5 Conclusions

This work proposed an approach to automatic conceptual taxonomy extraction from natural language texts. It works by mixing different techniques in order to identify relevant terms/concepts in the text, group them by similarity and generalize them to identify portions of a hierarchy. Preliminary experiments show that the approach can be viable, although extensions and refinements are needed to improve its effectiveness. In particular, a study on how to set standard suitable weights for concept relevance assessment is needed. A reliable outcome might help users in understanding the text content and machines to automatically performing some kind of reasoning on the resulting taxonomy.

References

[1] Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure trees. In *LREC*, 2006.

[2] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998.

[3] S. Ferilli, M. Biba, T.M. Basile, and F. Esposito. Combining qualitative and quantitative keyword extraction methods with document layout analysis. In *Post-proceedings of the 5th Italian Research Conference on Digital Library Management Systems (IRCDL-2009)*, pages 22–33, 2009.

[4] S. Ferilli, M. Biba, N. Di Mauro, T.M. Basile, and F. Esposito. Plugging taxonomic similarity in first-order logic horn clauses comparison. In *Emergent Perspectives in Artificial Intelligence*, Lecture Notes in Artificial Intelligence, pages 131–140. Springer, 2009.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The weka data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[6] R.W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160, 1950.

[7] Dan Klein and Christopher D. Manning. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2003.

[8] Sang Ok Koo, Soo Yeon Lim, and Sang-Jo Lee. Constructing an ontology based on hub words. In *ISMIS'03*, pages 93–97, 2003.

[9] A. Maedche and S. Staab. Mining ontologies from text. In *EKAW*, pages 189–202, 2000.

[10] A. Maedche and S. Staab. The text-to-onto ontology learning environment. In *ICCS-2000 - Eight International Conference on Conceptual Structures, Software Demonstration*, 2000.

[11] Bernardo Magnini and Gabriela Cavagli. Integrating subject field codes into wordnet. pages 1413–1418, 2000.

[12] Yutaka Matsuo and Mitsuru Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13:2004, 2003.

[13] N. Ogata. A formal ontology discovery from web documents. In *Web Intelligence: Research and Development, First Asia-Pacific Conference (WI 2001)*, number 2198 in Lecture Notes on Artificial Intelligence, pages 514–519. Springer-Verlag, 2001.

[14] Alessandro Cucchiarelli Paola Velardi, Roberto Navigli and Francesca Neri. Evaluation of OntoLearn, a methodology for automatic population of domain ontologies. In Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini, editors, *Ontology Learning from Text: Methods, Applications and Evaluation.* IOS Press, 2006.

[15] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, November 1975.

[16] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138, Morristown, NJ, USA, 1994. Association for Computational Linguistics.