The ContrastMedium Algorithm: Taxonomy Induction From Noisy Knowledge Graphs With Just a Few Links

Stefano Faralli[1], **Alexander Panchenko**[2], **Chris Biemann**[2] **and Simone Paolo Ponzetto**[1]

[1]Data and Web Science Group, University of Mannheim, Germany
[2]Language Technology Group, University of Hamburg, Germany
{stefano,simone}@informatik.uni-mannheim.de
{panchenko,biemann}@informatik.uni-hamburg.de

Abstract

In this paper, we present ContrastMedium, an algorithm that transforms noisy semantic networks into full-fledged, clean taxonomies. ContrastMedium is able to identify the embedded taxonomy structure from a noisy knowledge graph without explicit human supervision such as, for instance, a set of manually selected input root and leaf concepts. This is achieved by leveraging structural information from a companion reference taxonomy, to which the input knowledge graph is linked (either automatically or manually). When used in conjunction with methods for hypernym acquisition and knowledge base linking, our methodology provides a complete solution for end-to-end taxonomy induction. We conduct experiments using automatically acquired knowledge graphs, as well as a SemEval benchmark, and show that our method is able to achieve high performance on the task of taxonomy induction.

1 Introduction

Recent years have witnessed an impressive amount of work on automatic construction of wide-coverage knowledge resources. Web-scale open information extraction systems like NELL (Carlson et al., 2010) or ReVerb (Fader et al., 2011) have been successful in acquiring massive amounts of machine-readable knowledge by effectively tapping large amounts of text from Web pages. However, the output of these systems does not consist of a clean, fully-semantified output. Such output, on the other hand, could be provided by the vocabulary of large-scale ontologies like DBpedia (Bizer et al., 2009) or YAGO (Hoffart et al., 2013) and the integration of open and closed information extraction approaches (Dutta et al., 2014). The use of an encyclopedia-centric (e.g., Wikipedia-based) dictionary of entities leads to poor coverage of domain-specific terminologies (Faralli and Navigli, 2013). This can be alleviated by constructing knowledge bases of ever increasing coverage and complexity from the Web (Wu et al., 2012; Gupta et al., 2014; Dong et al., 2014) or by community efforts (Bollacker et al., 2008). However, the focus on large size and wide coverage at entity level has led all these resources to avoid the complementary problem of curating and maintaining a clean taxonomic backbone with as minimal supervision as possible. That is, no resource, to date, integrates structured information from existing wide-coverage knowledge graphs with empirical evidence from text for the explicit goal of building full-fledged taxonomies consisting of a clean and fully-connected directed acyclic graph (DAG). This is despite the fact that taxonomies have been known for a long time to provide valid tools to represent domain-specific knowledge with dozens of scientific, industrial and social applications (Glass and Vessey, 1995).

In taxonomy induction, the required domain knowledge can be acquired with many different methods for hypernym extraction, ranging from simple lexical patterns (Hearst, 1992; Oakes, 2005; Kozareva and Hovy, 2010) to statistical and machine learning techniques (Caraballo, 1999; Agirre et al., 2000; Ritter et al., 2009; Velardi et al., 2013). Recent efforts, such as Microsoft's Probase (Wu et al., 2012) or the WebIsaDB (Seitner et al., 2016) similarly focus on 'local' extraction of single hypernym relations, and do not address the problem of how to combine these single relations into a coherent taxonomy. When taxonomies are automatically acquired, their cleaning (also called "pruning") becomes a mandatory step (Velardi et al., 2013).

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 590–600,
Valencia, Spain, April 3-7, 2017.

The contributions of this paper are two-fold:

1. We introduce a new algorithm, named ContrastMedium, which, given a noisy knowledge graph and its (possibly automatically generated) links to a companion taxonomy, is able to output a full-fledged taxonomy. Information from the reference taxonomy is projected onto the input noisy graph to automatically acquire topological clues, which are then used to drive the cleaning process. The reference taxonomy provides us with ground-truth taxonomic relations that make our knowledge-based method not truly unsupervised *sensu stricto*. However, the availability of resources like, for instance, WordNet (Fellbaum, 1998) or BabelNet (Navigli and Ponzetto, 2012) implies that these requirements are trivially satisfied;

2. We combine our approach with an unsupervised framework for knowledge acquisition from text (Faralli et al., 2016) to provide a full end-to-end pipeline for taxonomy induction from scratch.

2 Related Work

Knowledge Bases (KBs) can be created in many different ways depending on the availability of external resources and specific application needs. Recently, much work in Natural Language Processing focused on Knowledge Base Completion (Nickel et al., 2016a, KBC), the task of enriching and refining existing KBs. Many different methods have been explored for KBC, including exploitation of resources such as text corpora (Snow et al., 2006; Mintz et al., 2009; Aprosio et al., 2013) or other KBs (Wang et al., 2012; Bryl and Bizer, 2014) for acquiring additional knowledge. Alternative approaches, in contrast, primarily rely on existing information from the KB itself (Socher et al., 2013; Nickel et al., 2016b) used as ground-truth to simultaneously learn continuous representations of KB concepts and relations, which are used to infer additional KB relations. Finally, Open Information Extraction methods looked at ways to extract large amounts of facts from Web-scale corpora in order to acquire open-domain KBs (Etzioni et al., 2011; Faruqui and Kumar, 2015, *inter alia*);

In this paper, we focus on a different, yet complementary task, which is a necessary step when inducing novel KBs from scratch, namely extracting clean taxonomies from noisy knowledge graphs. State-of-the-art algorithms differ by the amount of human supervision required and their ability to respect some topological properties while pruning. Approaches like those of Kozareva and Hovy (2010), Velardi et al. (2013) and Kapanipathi et al. (2014), for instance, apply different topological pruning strategies that require to specify the root and leaf concept nodes of the KB in advance – i.e., a predefined set of abstract top-level concepts and lower terminological nodes, respectively. The approach of Faralli et al. (2015) avoids such supervision on the basis of an iterative method that uses an efficient variant of topological sorting (Tarjan, 1972) for cycle pruning. Such lack of supervision, however, comes at the cost of not being able to preserve the original connectivity between the top (abstract) and the bottom (instance) concepts. Random edge removal (Faralli et al., 2015), in fact, can lead to disconnected components, a problem shared with the OntoLearn Reloaded approach (Velardi et al., 2013), which cannot ensure such property when used to approximate the solution for a large noisy graph.

Our work goes one step beyond the previous contributions by presenting a new efficient algorithm that is able to extract a clean taxonomy from a noisy knowledge graph without needing to know in advance – that is, having to manually specify – the top-level and leaf concepts of the taxonomy, while preserving the overall connectivity of the graph. We achieve this by projecting the information from a reference KB such as, for instance, WordNet (Fellbaum, 1998), onto the input noisy KB on the basis of pre-existing KB links – which in turn can be automatically generated with high precision using any of the existing solutions for KB mapping (Navigli and Ponzetto, 2012; Faralli et al., 2016, *inter alia*) or by relying on ground truth information from the Linguistic Linked Open Data cloud (Chiarcos et al., 2012).

Some aspects of the proposed approach – namely, the propagation of the nodes' weights through the graph, which we metaphorically represent as the flow of a contrast medium across nodes (Section 3.3) – are somewhat similar in spirit to spreading activation (Collins and Loftus, 1975) and random walks on graphs (Lovász, 1993) approaches. However, in contrast to spreading activation approaches we leverage the graph directionality in order to reach all the possible nodes within the same connected components. More-

over, in contrast to random walks on graphs our method is deterministic in nature. Here, we argue for the choice of a deterministic approach, like ours, that does not require tuning of parameters: its termination is guaranteed by the number of iterations, which we bind by the maximal diameter $|E|$ for a graph $G = (V, E)$. Generally, random walk algorithms would provide an approximation that may lead to a less precise estimation of the order induced by the contrast medium level.

3 The ContrastMedium Algorithm

3.1 Problem Statement

Our work builds upon the notion of a **noisy knowledge graph** (NKG), which consists of a directed graph $G = (V, E)$ where V is a set of concepts and E the set of labelled binary semantic relations – e.g., those found between synsets like, for instance, hypernymy or meronymy within a semantic network like WordNet. In a NKG we assume both V and E to have been acquired automatically, e.g., in order to induce a domain-aware or a general purpose knowledge base. Additionally, we consider for our purposes the **hypernymy graph** $T = (T_V, T_E)$ of G, the subgraph made up of the hypernymy (i.e., *isa*-labeled) edges of E. Since T is a subgraph of G, we can expect that the former inherits a certain amount of noise from the latter.

Noise within hypernymy graphs can be further classified into: i) *noisy nodes*, the concepts that do not belong to a specific target vocabulary, e.g., domain concepts for domain-specific KBs, such as `Jaguar Cars` within a zoological taxonomy; ii) *noisy edges*, the wrongly-acquired relations between unrelated concepts or out-of-domain relations, e.g., `Jaguar Cars` *isa* `Feline`; iii) *cycles of hypernymy relations*, such as those derived from counts over very large corpora (Seitner et al., 2016), e.g., `jaguar` (`Panthera onca`) → `feline` → `animal` → `jaguar` (`Panthera onca`). We accordingly define the task of extracting a clean taxonomy from a NKG as that of pruning the cycles, as well as the noisy edges and nodes, from the hypernymy subgraph T of G.

3.2 Resources Used

In order to enable end-to-end taxonomy induction from scratch, we combine our general approach with existing KBs that have been automatically induced from text and linked to reference lexical knowledge bases on the basis of unsupervised methods. To this end, we use the linked disambiguated distributional KBs from Faralli et al. (2016)[1], which are built in three steps:

1) **Learning a JoBimText model.** Initially, a sense inventory is created from a large text collection using the pipeline of the JoBimText project (Biemann and Riedl, 2013).[2] The resulting structure contains disambiguated proto-concepts (i.e., senses), their similar and related terms, as well as aggregated contextual clues per proto-concept.

2) **Disambiguation of related terms.** Similar terms and hypernyms associated with a proto-concept are fully disambiguated based on the partial disambiguation from step (1). The result is a proto-conceptualization (PCZ), where all terms have a sense identifier.

3) **Linking to a lexical resource.** The PCZ is automatically aligned with an existing lexical resource (LR) such as WordNet or BabelNet. For example, `bridge:NN:3` is linked to the Babel synset `bn:00013077n` (the 'infrastructure' sense). That is, a mapping between the two sense inventories is created to combine them into a new extended sense inventory, a *hybrid aligned resource*.

Table 1 shows the proto-conceptualization entries for the polysemous terms *bridge* and *link*, namely their figurative ("bridge:NN:2" and "link:NN:1") and concrete 'infrastructure' ("bridge:NN:3" and "link:NN:0") senses, respectively. JoBimText models provide sense distinctions that are only partially disambiguated: the list of similar and hypernyms terms of each sense, in fact, does not carry sense information. Consequently, a semantic closure procedure is applied in order to obtain a PCZ and arrive at sense representation in which all terms get assigned a unique, best-fitting sense identifier (see Faralli et al. (2016) for details).

PCZs consist of a rich, yet noisy, disambiguated semantic network automatically induced from large amounts of text: links to existing lexical resources provide us a source of external supervision that can be leveraged to clean them and turn them into full-fledged taxonomies. Steps 1–3 are unsupervised by nature. Consequently, when

[1] `https://madata.bib.uni-mannheim.de/171/`
[2] `http://www.jobimtext.org`

entry	similar terms	hypernyms
bridge:NN:2	gap:NN:2, divide:NN:2, link:NN:1, ...	issue:NN:2, topic:NN:3, ...
bridge:NN:3	road:NN:0, highway:NN:1, overpass:NN:3 ...	infrastructure:NN:1, project:NN:1, ...
link:NN:0	connection:NN:3, correlation:NN:1, linkage:NN:1 ...	service:NN:6, feature:NN:0, ...
link:NN:1	relationship:NN:1, interaction:NN:1, divide:NN:0 ...	problem:NN:1, topic:NN:3 ...

Table 1: Excerpt of a proto-conceptualization (PCZ) for the words "bridge:NN" and "link:NN".

combined with our algorithm they provide a complete framework for fully unsupervised taxonomy induction from scratch. Note, however, that our approach offers a general solution to the problem of taxonomy cleaning. In an additional set of experiments, we apply it to different automatically generated taxonomies from a SemEval task in a more controlled setting where we rely on a few manually created KB links only.

3.3 The ContrastMedium Algorithm

At its core, our algorithm relies on the notion of a **linked noisy knowledge graph** (LNKG). This consists of a quintuple (G, KB, KB_{root}, λ, M) where: i) $G = (V_G, E_G)$ is a noisy knowledge graph; ii) $KB = (V_{KB}, E_{KB})$ is a companion knowledge base providing a ground-truth taxonomy; iii) KB_{root} is the root node of the reference knowledge base KB (if several top-level nodes exist, an artificial root can be created by connecting them all); iv) λ is a conventional symbol to represent the "undefined concept", i.e., a place-holder for empty mappings; v) $M : V_G \rightarrow V_{KB} \cup \{\lambda\}$ is the function, which maps nodes of V_G into nodes of V_{KB} or into the undefined concept λ. The key ideas behind ContrastMedium are:

- Identification of important topological clues from the companion knowledge base KB in order to hierarchically sort the concepts in G. For our purposes, KB is expected to be able to provide ground-truth taxonomic relations that can be safely projected onto G to guide the cleaning process: that is, we assume it to be reasonably clean. In contrast, we do not make any assumption on how KB has been created: our approach can be used with either manually created taxonomies like WordNet or (semi-)automatically induced ones, provided they are of sufficient quality. Hence, our method is knowledge-based without the need of further supervision other than that contained in KB;

- Projection of topological clues from KB back onto the LNKG G on the basis of the links found in the mapping M. Similarly to the case of the reference knowledge base, we do not make any assumption on how the links between G and KB have been created: while there exists different methods to automatically link (lexical) knowledge bases (Navigli and Ponzetto, 2012; Faralli et al., 2016), we later show that it is also possible to achieve state-of-the-art performance with a few manually given links;

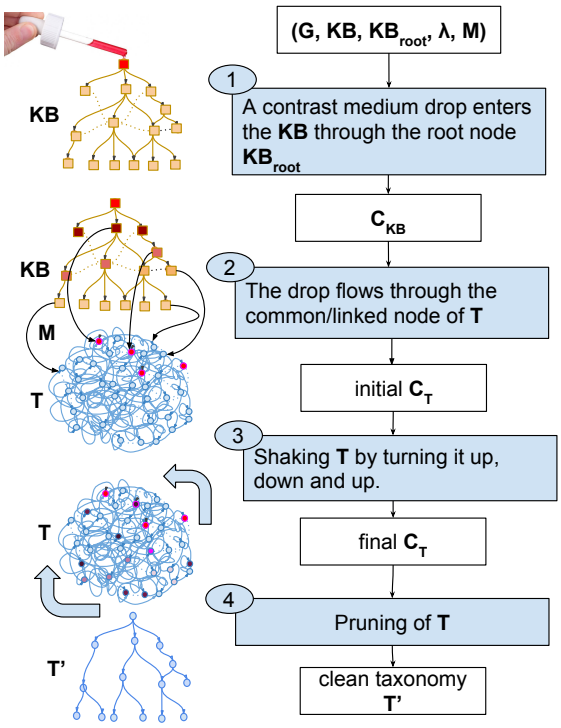

Figure 1: ContrastMedium: algorithm workflow.

- Propagation of the topological clues across the entire NKG G. That is, to cope with the partial coverage of automatic mappings, as well as the need to reduce the number of manually created KB links, we apply a signal propagation technique that solely relies on the structure of G;

- To make use of the resulting topological clues to drive the taxonomy pruning process. That is, propagated topological clues from KB are additionally leveraged to ensure that the output

ALGORITHM 1: The ContrastMedium algorithm.

Input: $(G = (V, E), KB = (V_{KB}, E_{KB}), KB_{root}, \lambda, M)$
Output: hypernymy graph T' of G, s.t. T' has no cycles.

// Estimating clues from KB (Fig. 1, step 1)
1 $\forall x \in V_{KB}$: $C_{KB}(x) = 0$;
2 injectContrastMedium(KB, KB_{root});
// Transferring clues from KB to G (Fig. 1, step 2)
3 $T = (V_T, E_T) \longleftarrow$ hypernymyGraph(G);
4 $\forall x \in V_T : C_T(x) = 0$;
5 transferClues(M, KB, T, C_{KB}, C_T);
// Shaking the graph T (Fig. 1, step 3)
6 shake(UP, T, C_T); // propagate through in-edges
7 shake(DOWN, T, C_T); // propagate through out-edges
8 shake(UP, T, C_T); // propagate through in-edges
// Pruning the graph T (Fig. 1, step 4)
9 T' = prune(T, C_T);
10 return T';

ALGORITHM 2: The Shake routine.

Input: $direction$ may be UP or DOWN,
$graph = (V_{graph}, E_{graph}), C_{graph}$
Output: the updated C_{graph}

1 **foreach** $x \in V_{graph}$ **do**
2 $\quad Current_{graph}(x) = C_{graph}(x); Flown_{graph} = 0.0;$
// iteratively propagates the CM
3 **for** $i = 0$ *to* $|E_{graph}| - 1$ **do**
4 $\quad$ **foreach** $x \in V_{graph}$ **do**
5 $\quad\quad InOut_{graph}(x) = 0.0;$
6 $\quad$ **foreach** x *s.t.* $Current_{graph}(x) > 0.0$ **do**
7 $\quad\quad CMlevel = Current_{graph}(x);$
8 $\quad\quad$ **if** $direction ==$ *DOWN* **then**
9 $\quad\quad\quad O = outgoingEdges(x, graph);$
10 $\quad\quad\quad$ **foreach** $(x, y) \in I$ **do**
11 $\quad\quad\quad\quad InOut_{graph}(y) += \frac{CMlevel}{max(|O|,1)};$
12 $\quad\quad$ **else**
13 $\quad\quad\quad I = incomingEdges(x, graph);$
14 $\quad\quad\quad$ **foreach** $(y, x) \in I$ **do**
15 $\quad\quad\quad\quad InOut_{graph}(y) += \frac{CMlevel}{max(|I|,1)};$
16 $\quad\quad Flown_{graph}(x) += CMlevel;$
17 $\quad$ **foreach** $x \in V_{graph}$ **do**
18 $\quad\quad Current_{graph}(x) = InOut_{graph}(x);$
19 **foreach** $x \in V_{graph}$ **do**
20 $\quad C_{graph}(x) = Flown_{graph}(x);$

results in a proper taxonomic structure.

We rely on the metaphor of a contrast medium (CM) to describe our approach, which is summarized in Figure 1. In the context of clinical analysis, a CM is injected into the human body to highlight specific complex internal body structures (in general, the venuous system). In a similar fashion, we detect the topological structure of a graph by propagating a certain amount of CM that we initially inject through the node KB_{root} of the companion knowledge base KB. The highlighted structure indicates the distance of a node with respect to the node KB_{root}. Then the lowest values of contrast medium indicate the leaf terminological nodes. The observed quantities are then transferred to corresponding nodes of the noisy graph by the mapping M. Next, the medium is propagated by 'shaking' the noisy graph. We let the fluid reach all the components G by alternating two phases of propagation: letting the CM to flow through both incoming ('shake up'); and outgoing ('shake down') edges. At the end, we use the partial order induced by the observed node level of CM to drive the pruning phase, and 'stretch' the original NKG G into a proper DAG.

Our approach is presented in Algorithms 1 and 2.[3] It consists of the following main steps:

1) CM injection Cf. Figure 1, block 1 and Algorithm 1, lines 1-2. We initially define the function $C_{KB} : V_{KB} \rightarrow [0.0 - 1.0]$ and assign a zero contrast medium level to all the nodes of the KB graph $C_{KB}(x) = 0, x \in V_{KB}$ (line 1). Next, we call the routine 'injectContrastMedium' which: 1) assigns an initial contrast level equals to 1.0 to the node KB_{root} of the KB graph; ii) uses the routine "Shake" with the direction parameter equals to "DOWN" (see Algorithm 2 and Step 3 "Graph shaking" for more details) to let the CM drop through KB. In practice, the shaking routine implements a node contrast medium level propagation algorithm following the outgoing ('down') or the incoming ('up') edges of the graph.

2) CM transfer Cf. Figure 1, block 2 and Algorithm 1, lines 3-5. In the next phase, we first extract the hypernymy subgraph $T = (V_T, E_T)$ of G (see Section 3.1) and then follow the links in the mapping M to transfer the contrast medium levels, i.e., $C_T(y) = C_{KB}(x)$ (s.t. $x \in V_{KB}, y \in V_T, (y \rightarrow x) \in M$).

3) Graph shaking Cf. Figure 1, block 3 and Algorithm 1, lines 6-8. After having transferred the CM to the target hypernym graph T of G, we shake T to let the CM flow by traversing the incoming, the outgoing, and finally the incoming edges again – see Algorithm 2 for details on the 'Shake' routine. Note that these two kinds of propagation are needed since the CM needs to be propagated through all the nodes of the graph to highlight the topological clues we are searching for. In particular, in Algorithm 2 at each iteration t for each node $x \in V_{graph}$, depending on

[3] A demo is available at http://web.informatik.uni-mannheim.de/faralli/cm.html with examples of the application of ContrastMedium to a few simple LNKGs.

the value of the parameter $direction$ (line 8 and line 12): i) we observe a CM level for the node x (line 7); ii) if $direction ==$ DOWN (lines 9-11) we traverse all the outgoing edges (x, y) of x and propagate the observed CM level of x, otherwise ($direction ==$ UP, lines 13-15) we traverse the incoming edges (y, x) and propagate the CM level to the nodes y; iii) the value of $Flown_{graph}(x)$ is incremented by the observed CM level (line 16); iv) for each node x we reset the current observed value of the CM level with the portion of the liquid which has flown from the incoming or the outgoing edges during the propagation (lines 17-18).

Depending on the propagation direction, we have two different behaviours for the CM. When exiting a node x through out the outgoing edges ($direction ==$ DOWN) we increment the level of contrast medium of the reached nodes by the observed value of x divided by number of outgoing edges of x. By converse, when we climb ($direction ==$ UP) across the incoming edges of a node x we increment the CM level of the reached node by the observed CM quantity of x divided by the number of incoming edges of x.

Note that the sequence UP/DOWN/UP and the specular DOWN/UP/DOWN are the only ones from the 8 possible combinations which can guarantee the contrast medium to flow on the entire graph. We simply selected the first sequence since the final rank places candidate root nodes on the top (and candidate leaf nodes on the bottom).

4) Pruning Cf. Figure 1, block 4 and Algorithm 1, lines 9. Finally, we create a clean taxonomy T' by pruning the graph T on the basis of the contrast levels found in C_T. CM levels in C_T can be used to induce a order of the nodes that, intuitively, captures the level of conceptual abstraction for the nodes in T. We use them to produce a clean taxonomy as follows. We first sort the nodes $v \in V_T$ in a list $S = s_0, s_1, \ldots, s_{|V_T|-1}$ by the decreasing resulting CM level value in C_T. The nodes with a higher level of contrast medium are candidates to be at the top level while the ones at the end of the list are candidates to be leaf nodes of the output taxonomy. Next, the pruning routine starts from a graph $T' = (V_{T'} = V_T, E_{T'} = \emptyset)$ and for each node $s \in S$ (from the last node to the first) add to $E_{T'}$ all the edges of the kind $e = (y, s)$ where a path from y to s does not exists in T and with y belonging to one of the following: i) the set of peers $\{x \in S \text{ s.t. } C_T(x) = C_T(s)\}$; ii) the ascending ordered list of preceding ($x \in S$ s.t. $C_T(x) > C_T(s)$); iii) the ascending ordered list of following ($x \in S$ s.t. $C_T(x) < C_T(s)$)

Complexity analysis. The propagation step (Figure 1, blocks 1 and 3; Algorithm 2) costs $O(|E| * |V|)$ since we iteratively analyze all the nodes of V for a number $|E|$ of iterations. The final step of pruning (Figure 1, block 4), instead, can have a time cost $O(|V|^2 * (|E| + |V|))$, since, in the worst case, the algorithm must analyse all the possible pairs of vertices, and then test the existence of a directed path between the candidate pairs of nodes.

4 Experiments

We perform two sets of experiments. We first evaluate our approach when applied to large, automatically induced noisy knowledge graphs (Section 4.1) and then quantify the impact it can have to further improve the quality of the output of state-of-the-art taxonomy induction systems (Section 4.2).

4.1 Experiment 1: Pruning existing LNKG

We first apply ContrastMedium to a variety of knowledge graphs that have been automatically acquired and linked to reference KBs like WordNet and BabelNet using unsupervised methods (Section 3.2). Our research questions (RQs) are:

RQ1 Can we use ContrastMedium as component of a complete framework for fully unsupervised taxonomy induction from scratch?

RQ2 What is the quality of the resulting taxonomies?

4.1.1 Experimental Setting

We apply our pruning algorithm to the automatically acquired KBs presented by Faralli et al. (2016). These noisy knowledge graphs have been induced from large text corpora and include both taxonomic and other (i.e., related, topically associative) semantic relations (cf. Table 1), as well as automatically induced mappings to lexical knowledge bases like WordNet and BabelNet. These NKGs have been induced from a 100 million sentence news corpus (*news*) and from a 35 million sentence Wikipedia corpus (*wiki*), using different parameter values to generate sense inventories of different granularities (e.g., 1.8 vs. 6.0 average senses per term for the wiki-p1.8 and wiki-p6.0 datasets, respectively). Table 2 shows some of

	senses	polysemy		hypernyms		links	hypernymy graph	
dataset	#	avg.	max	#	avg.	#	$\|V_T\|$	$\|E_T\|$
news-p1.6	332k	1.6	18	15k	6.9	60k	170k	1.538k
news-p2.3	461k	2.3	17	15k	5.8	95k	225k	1.871k
wiki-p1.8	368k	1.8	15	15k	4.4	67k	185k	1.167k
wiki-p6.0	1.5M	6.0	36	52k	1.7	279k	394k	1.901k

Table 2: Dimensions of the four datasets adopted as linked noisy knowledge graphs (Faralli et al., 2016).

the dimensions for each of the four NKGs – number of senses, average and maximum sense polysemy, number and average hypernyms per sense, the number of linked senses to WordNet concepts (i.e., "links"), and the number of nodes and edges for the corresponding hypernymy graph. Since our algorithm primarily focuses on conceptual hierarchical (taxonomic) structures – referred to as the TBox in Knowledge Representation – we use the WordNet mappings only, since the manual inspection of the BabelNet mappings revealed that they are focused primarily on instances (that is, ABox statements). In order to have a complete quintuple for each NKG, we selected, for the companion KB, the top KB_{root} concept `entity` of the WordNet taxonomy (SynsetID `SID-00001740-N`).

4.1.2 Measures

We benchmark ContrastMedium using a variety of metrics that are meant to capture structural properties of the output taxonomies (to describe the impact of pruning on the original NKGs), as well as an estimation of their overall quality.

Edge compression: the ratio of the number of pruned edges over the total number of edges:

$$C_{E_{G,G'}} = \frac{|E_G| - |E_{G'}|}{|E_G|}$$

where E_G and $E_{G'}$ represent the number of edges found within the input (G) and pruned (G') taxonomy, respectively.

Pruning accuracy: the performance on a 3-way classification task to automatically detect the level of granularity of a concept as a proxy to quantify the overall quality of the output taxonomies. Pruning accuracy is estimated using gold-standard annotations that are created from a random sample of 1,000 nodes for each NKG. Two annotators with previous experience in knowledge acquisition and engineering were asked to provide for each concept whether it can be classified as: i) a root, top-level abstract concept – i.e., any of `entity`, `object`, etc. and more in general nodes that correspond to abstract concepts that we can expect to be part of a core ontology such as, for instance, DOLCE (Gangemi et al., 2002); ii) a leaf terminological node (i.e., instances such as `Lady Gaga` or `Porsche 911`); iii) or a middle-level concept (e.g., `celebrity` or `cars`, concepts not fitting into any of the previous classes). An adjudication procedure was used to resolve any discrepancy between the two annotators: the inter-annotator agreement after adjudication is $\kappa = 0.657$ (Fleiss, 1971), with most disagreement occurring on the identification of abstract, core ontology concepts.

A local 3-way classification task provides a rather crude way to estimate the performance on inducing hierarchical structures like taxonomies. Here, we use it primarily to benchmark how well ContrastMedium compares against other, structure-agnostic algorithms used within state-of-the-art solutions such as, for instance, Tarjan's topological sorting (Section 2), which only break cycles in a random fashion.

Given ground-truth concept granularity judgements, we compute standard accuracy for each of the three classes. That is, we compare the system outputs against the gold standards and obtain three accuracy measures: one for the root nodes (A_R), one for the nodes 'in the middle' (A_M) and finally one for the leaf nodes (A_L). For example a true positive root node is a node annotated as a root node in the gold standard and having no incoming edges in the pruned graph.

Error Reduction (ER): finally, we compute the relative error reduction of ContrastMedium against other, baseline approaches as:

$$\frac{Baseline_{errors}/|sample| - CM_{errors}/|sample|}{Baseline_{errors}/|sample|}$$

As *baseline* we use the approach of Faralli et al.

	Pruned Knowledge Graph						Pruning accuracy						ER
	ContrastMedium			Tarjan (baseline)			ContrastMedium			Tarjan (baseline)			
dataset	$\|V_{G'}\|$	$\|E_{G'}\|$	$C_{E_{G,G'}}$	$\|V_{G'}\|$	$\|E_{G'}\|$	$C_{E_{G,G'}}$	A_R	A_M	A_L	A_R	A_M	A_L	
news-p1.6	170k	1.536k	0.15%	170k	1.535k	0.18%	**98.9**	**98.3**	**99.3**	93.3	94.6	95.3	0.62
news-p2.3	225k	1.867k	0.19%	225k	1.866k	0.23%	**98.7**	**98.7**	**99.9**	95.7	94.7	95.6	0.50
wiki-p1.8	183k	1.165k	0.18%	183k	1.164k	0.22%	**97.6**	**94.7**	**97.3**	93.1	87.3	94.1	0.41
wiki-p6.0	394k	1.897k	0.18%	394k	1.896k	0.21%	**95.9**	**94.3**	**98.3**	89.5	90.1	92.8	0.50

Table 3: Structural analysis, pruning accuracies and error reduction (ER) for the four LNKGs.

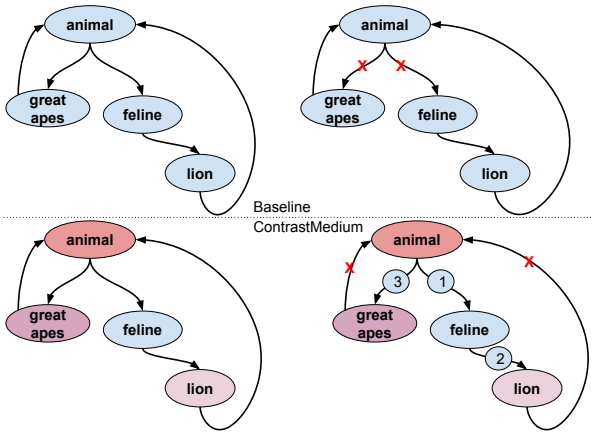

Figure 2: An example noisy graph and the different solutions provided by ContrastMedium and the baseline.

(2015) based on Tarjan's topological sorting (Section 2), which iteratively searches for a cycle (until no cycle can be found) and randomly removes an edge from it. To the best of our knowledge, this is the only algorithm that we can fairly compare with, since alternative solutions all need to know the sets of root and leaf nodes in advance.

4.1.3 Results and discussion

Table 3 summarizes the performance of ContrastMedium on the four automatically acquired NKGs. The results show that the pruning impact of our approach is lower than that of the baseline (an average of 1K edges of difference, cf. columns 3 and 6), which also determines higher edge compression $C_{E_{G,G'}}$ values for the baseline method. Despite being less aggressive in terms of the number of edges pruned, ContrastMedium outperforms the Tarjan-based algorithm on all datasets in terms of accuracy. Thanks to our method, in fact, we are able to achieve, even despite the baseline already reaching very high performance levels (well above 90% accuracy), improvements of up to 6 points, with an overall error reduction between around 40% and 60%. To provide an intuition of why ContrastMedium clearly outperforms the baseline approach, we provide in Figure 2 an exemplified depiction of a typical case on which the baseline fails (based on a manually inspected random sample). In our example, the Tarjan baseline first detects the cycle $C_1 = (\texttt{lion} \rightarrow \texttt{animal} \rightarrow \texttt{feline} \rightarrow \texttt{lion})$ and randomly decides to break it by removing the edge (`animal` → `feline`). Next, it detects the cycle $C_2 = (\texttt{animal} \rightarrow \texttt{great apes} \rightarrow \texttt{animal})$ and randomly decides to break it by removing the edge (`animal` → `great apes`). ContrastMedium, instead, after the shaking of the graph can leverage the partial ordering of the nodes (based on the concept granularity of the corresponding concepts) to select the edges (`animal`, `feline`), (`feline`, `lion`) and (`animal`, `great apes`), while removing all remaining wrong and redundant edges.

4.2 Experiment 2: SemEval-15 task 17

We next evaluate the overall impact of our approach within an existing benchmark for the taxonomy induction task. Intuitively, most of the benefits from our method derive from the "gold standard" information of the companion KB, and its linking to the NKG, which act as a source of supervision. Consequently, we address the research question of how much (pseudo-)supervision our method needs in terms of KB links, and whether it can be used to improve the state-of-the-art on the task of taxonomy induction.

4.2.1 Experimental Setting

We use the benchmark data from the SemEval-15 task 17 "Taxonomy Extraction Evaluation: TExEval" (Bordea et al., 2015), since it provides us with gold-standard datasets and system outputs within a standard, easy-to-reproduce setting. Initially, we select from the participating systems[4] the two best performing taxonomies based on the Cumula-

[4]Cf. Table "Comparative Evaluation" at `http://alt.qcri.org/semeval2015/task17/index.php?id=evaluation`

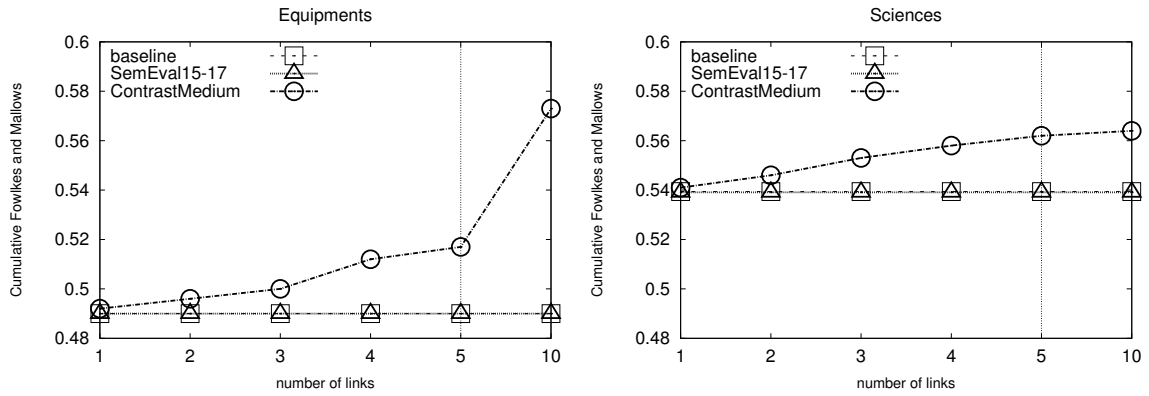

Figure 3: Performance on the SemEval-15 TExEval dataset (Cumulative Fowlkes&Mallows measure).

tive Fowlkes&Mallows (CF&M) measure (Velardi et al., 2012), the *Equipments* and *Sciences* taxonomies from the INRIASAC and the LT3 teams respectively. We next apply our approach to these taxonomies, in order to clean them in a post-processing fashion. By selecting the top-systems we can see how far we can advance the state-of-the-art overall. Besides, these two taxonomies are also the ones containing the highest number of cycles, giving the application of our cleaning algorithm a more challenging (and meaningful) setting. To remove the effects of automatic linking and quantify the amount of manual efforts needed by our approach, 10 random concepts from each of these resources are manually linked to WordNet, and the taxonomies subsequently pruned using ContrastMedium and the baseline. We then evaluate performance following the task's experimental setting and compute the CF&M measure for different levels of manually-created KB links.

4.2.2 Results and discussion

In Table 3, we report the performance on the SemEval task for the two selected input taxonomies. Results on the structural similarities of the pruned taxonomies with the gold standard ones, computed using the CF&M measure, indicate that, thanks to ContrastMedium and with a minimal human effort – the creation of just a few KB links (up to 10), which are needed only when automatic linking is not available – it is possible to boost the quality of taxonomies using state-of-art methods by a large margin. For instance, in the case of the *Equipments* taxonomy, we improve up to 7 points. The baseline, which only breaks cycles, is not able to reassess the graph structure and only provides very small improvements to the submitted NKGs. Overall, the results show that ContrastMedium leads to competitive performance on a hard, realistic benchmark such as TExEval, achieving the best overall results for both taxonomies. That is, our algorithm is able to improve the state-of-the-art on taxonomy induction by additionally boosting the quality of existing top-performing systems for this task: this is achieved on the basis of a minimally supervised approach that only requires a few links to a reference KB, which is used to provide ground-truth taxonomic relations and guide the cleaning process.

5 Conclusions

In this paper, we presented ContrastMedium, a novel algorithm that can be applied to automatically linked noisy knowledge graphs to provide an end-to-end solution for fully unsupervised taxonomy induction from scratch, i.e., without any human effort. Our results indicate that ContrastMedium can be successfully applied to a wide range of automatically acquired KBs, ranging from large linked noisy knowledge graphs all the way to small-scale induced taxonomies to produce high-quality *isa* hierarchies that achieve state-of-the-art results on SemEval benchmarks. As future work, we plan to improve the scalability of the algorithm, in particular its time complexity order, and apply it to Web-scale resources like the WebIsaDB (Seitner et al., 2016) or state-of-the-art approaches like TAXI (Panchenko et al., 2016), as well as to publicly release the created resources.

Acknowledgments

We acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) under the JOIN-T project.

References

Eneko Agirre, Olatz Ansa, Eduard H. Hovy, and David Martínez. 2000. Enriching very large ontologies using the WWW. In *Proceedings of the ECAI 2000 Workshop on Ontology Learning*.

Alessio P. Aprosio, Claudio Giuliano, and Alberto Lavelli. 2013. Extending the coverage of DBpedia properties using distant supervision over Wikipedia. In *Proceedings of the NLP & DBpedia workshop co-located with the 12th International Semantic Web Conference (ISWC 2013)*, pages 20–31.

Chris Biemann and Martin Riedl. 2013. Text: Now in 2D! A Framework for Lexical Expansion with Contextual Similarity. *Journal of Language Modelling*, 1(1):55–95.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia – A crystallization point for the web of data. *Journal Web Semantics*, 7(3):154–165.

Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 1247–1250.

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. Semeval-2015 task 17: Taxonomy extraction evaluation (TExEval). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 902–910.

Volha Bryl and Christian Bizer. 2014. Learning conflict resolution strategies for cross-language wikipedia data fusion. In *Proceedings of the 23rd International World Wide Web Conference*, pages 1129–1134.

Sharon A. Caraballo. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 120–126.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pages 1306–1313.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2012. Linking linguistic resources: Examples from the Open Linguistics Working Group. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata*, pages 201–216. Springer.

Allan M. Collins and Elizabeth F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407 – 428.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 601–610.

Arnab Dutta, Christian Meilicke, and Simone Paolo Ponzetto. 2014. A probabilistic approach for integrating heterogeneous knowledge sources. In *Proceedings of the 11th Extended Semantic Web Conference*, pages 286–301.

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam. 2011. Open information extraction: The second generation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 3–10.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.

Stefano Faralli and Roberto Navigli. 2013. Growing Multi-Domain Glossaries from a Few Seeds using Probabilistic Topic Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 170–181.

Stefano Faralli, Giovanni Stilo, and Paola Velardi. 2015. Large scale homophily analysis in twitter using a twixonomy. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 2334–2340.

Stefano Faralli, Alexander Panchenko, Chris Biemann, and Simone P. Ponzetto. 2016. Linked disambiguated distributional semantic networks. In *The Semantic Web – ISWC 2016: 15th International Semantic Web Conference*, pages 56–64.

Manaal Faruqui and Shankar Kumar. 2015. Multilingual open relation extraction using cross-lingual projection. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1356.

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

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schneider. 2002. Sweetening ontologies with DOLCE. In *Proceedings of the 13th International Conference on*

Knowledge Engineering and Knowledge Management, pages 166–181.

Robert L. Glass and Iris Vessey. 1995. Contemporary application-domain taxonomies. *IEEE Software*, 12(4):63–76.

Rahul Gupta, Alon Y. Halevy, Xuezhi Wang, Steven Euijong Whang, and Fei Wu. 2014. Biperpedia: An ontology for search applications. In *Proceedings of the 40th International Conference on Very Large Data Bases*, pages 505–516.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, pages 539–545.

Johannes Hoffart, Fabian Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194(28):28–61.

Pavan Kapanipathi, Prateek Jain, Chitra Venkataramani, and Amit Sheth. 2014. User interests identification on Twitter using a hierarchical knowledge base. In *The Semantic Web: Trends and Challenges*, pages 99–113. Springer.

Zornitsa Kozareva and Eduard Hovy. 2010. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1110–1118.

László Lovász. 1993. Random walks on graphs: A survey. *Combinatorics*, 2:146.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016a. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso A. Poggio. 2016b. Holographic embeddings of knowledge graphs. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 1955–1961.

Michael P. Oakes. 2005. Using Hearst's rules for the automatic acquisition of hyponyms for mining a pharmaceutical corpus. In *Proceedings of the RANLP 2005 Text Mining Workshop*, pages 63–67.

Alexander Panchenko, Stefano Faralli, Eugen Ruppert, Steffen Remus, Hubert Naets, Cédrick Fairon, Simone Paolo Ponzetto, and Chris Biemann. 2016. TAXI at SemEval-2016 Task 13: A taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, pages 1320–1327.

Alan Ritter, Stephen Soderland, and Oren Etzioni. 2009. What is this, anyway: Automatic hypernym discovery. In *Proceedings of the AAAI Spring Symposium on Learning by Reading and Learning to Read*, pages 88–93.

Julian Seitner, Christian Bizer, Kai Eckert, Stefano Faralli, Robert Meusel, Heiko Paulheim, and Simone Paolo Ponzetto. 2016. A large database of hypernymy relations extracted from the web. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 360–367.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 801–808.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*, pages 926–934.

Robert Tarjan. 1972. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1:146–160.

Paola Velardi, Roberto Navigli, Stefano Faralli, and Juana María Ruiz-Martínez. 2012. A new method for evaluating automatically learned terminological taxonomies. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1498–1504.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. OntoLearn Reloaded: A Graph-Based Algorithm for Taxonomy Induction. *Computational Linguistics*, 39(3):665–707.

Zhichun Wang, Juanzi Li, Zhigang Wang, and Jie Tang. 2012. Cross-lingual knowledge linking across wiki knowledge bases. In *Proceedings of the 21st International World Wide Web Conference*, pages 459–468.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 481–492.