Leveraging and Balancing Heterogeneous Sources of Evidence in Ontology Learning

Gerhard Wohlgenannt

Vienna Univ. of Economics and Business, Welthandelsplatz 1, 1200 Wien, Austria
gerhard.wohlgenannt@wu.ac.at
http://www.wu.ac.at

Abstract. Ontology learning (OL) aims at the (semi-)automatic acquisition of ontologies from sources of evidence, typically domain text. Recently, there has been a trend towards the application of multiple and heterogeneous evidence sources in OL. Heterogeneous sources provide benefits, such as higher accuracy by exploiting redundancy across evidence sources, and including complementary information. When using evidence sources which are heterogeneous in quality, amount of data provided and type, then a number of questions arise, for example: How many sources are needed to see significant benefits from heterogeneity, what is an appropriate number of evidences per source, is balancing the number of evidences per source important, and to what degree can the integration of multiple sources overcome low quality input of individual sources? This research presents an extensive evaluation based on an existing OL system. It gives answers and insights on the research questions posed for the OL task of concept detection, and provides further hints from experience made. Among other things, our results suggest that a moderate number of evidences per source as well as a moderate number of sources resulting in a few thousand data instances are sufficient to exploit the benefits of heterogeneous evidence integration.

Keywords: heterogeneous evidence sources, ontology learning, evidence integration, spreading activation

1 Introduction

Ontologies are a cornerstone technology and backbone for the Semantic Web, but the manual creation of ontologies is cumbersome and expensive, therefore there have been many efforts towards (semi-)automatic ontology generation in order to assist ontology engineers.

The process of ontology learning (typically from text) in a first step extracts facts (lexical entries) and patterns (evidence) from text, and then turns them into shareable high-level constructs. This includes the identification of domain concepts, which is an ontology learning (OL) task building on term extraction and the detection of synonyms [2].

OL evolved from working on static domain text to Web sources, and more recently there are a few approaches that make use of multiple and heterogeneous data sources (see next section for more details). The introduction of heterogeneous sources into the learning process offers the potential for higher levels of accuracy, on the other hand there are challenges regarding the meaningful integration and balancing (of the impact) of sources. Manzano-Macho et al. [6] list some of the reasons for increased accuracy when using heterogeneous evidence sources: (i) redundancy of information in different sources represents a measure of relevance and trust, and (ii) additional sources can provide complementary data and valuable information that the other sources did not detect.

The question arising is to quantify the gains in accuracy in various OL tasks when using heterogeneous evidence sources. In this paper we take a detailed look on gains in the concept detection task. So, the research question is: How does the number and the characteristics of heterogeneous evidence sources affect accuracy (ie. the ratio of relevant concept candidates) in concept detection? In other words, the problem is as follows: We start with an OL system that includes a number of (heterogeneous) evidence acquisition methods, which basically provide terminology (heterogeneous lists of terms). These are the input, the output of concept detection are a number of domain concept candidates. In the evaluation section we study the impact of the (i) number of evidence sources, (ii) number of evidences per source, (iii) heterogeneity and quality of sources and (iv) the balance between sources on the accuracy of concept detection.

The evidence used in the OL system is heterogeneous in various respects. It originates from different sources such as Web documents, social Web APIs, and structured sources, and from different extraction methods applied. This leads to heterogeneity regarding the quality of evidence, the vocabulary used, the number of evidences and the dynamics of the source (see Section 4).

The experiments are conducted with an OL system (see Section 3) that generates lightweight ontologies using the spreading activation algorithm [5] to integrate evidence. Lightweight ontologies typically only contain concepts, taxonomic relations and unlabeled non-taxonomic relations, and are applied in many areas, e.g. to fuel everyday applications like Web search and enabling intelligent systems [19]. For the experiments, the architecture generated lightweight ontologies in two different domains ("climate change" and "tennis") in monthly intervals from scratch. As spreading activation is a simple and intuitive way to integrate heterogeneous evidence, the results can largely be generalized to other OL systems and integration logics for heterogeneous evidence which use a similar approach.

The outline of the paper is as follows: After presenting related work in Section 2, Section 3 introduces the OL system used in the experiments. Section 4 provides details about the heterogeneous sources of evidence. Results of the extensive experiments are found in Section 5, Section 6 concludes with a summary, the main contributions, and future work.

2 Related Work

Most OL systems learn ontologies from only one source, typically domain text, e.g. Text2Onto [4] or OntoLearn Reloaded [14]. Some authors, e.g. Sanchez and Moreno [12], combine corpus-based methods with Web statistics for ontology learning tasks. Others exploit structured data present in the current Semantic Web, e.g. Alani [1], who proposes a method for ontology building by cutting and pasting segments from online ontologies. More recently, some systems start to make use of heterogeneous evidence sources in OL. Using only one evidence source typically results in modest levels of accuracy [6], the combination of several sources may partially overcome this problem.

Manzano-Macho et al. [6] present an architecture which learns from multiple sources using a number of methods. In the acquisition layer the system learns hypotheses about candidate elements (the core terminology of the domain) which include a probability of relevance and relations to other candidate elements. Acquisition uses statistical methods as well as NLP tools and visual (HTML layout-based) methods. Furthermore, the system filters for domain relevance, detects domain concepts and taxonomic relations, and evaluates the resulting ontology against a pre-selected reference ontology. OntoElect [13] is methodology for ontology engineering, which applies term extraction to papers by domain experts. They also describe termhood saturation experienced when extending the collection of papers. Among the few papers which focus on OL from heterogeneous sources is also an approach by Cimiano et al. [3] to learn taxonomic relations. This method converts evidence into first order logic features, and then uses standard classifiers (supervised machine learning) on the integrated data to find good combinations of input sources. The input sources include data from lexico-syntactical pattern matching, head matching and subsumption heuristics applied to domain text. Völker et al. [15] propose a similar approach which uses the confidence scores of several heterogeneous methods as features in a classifier, aiming to enrich existing ontologies with disjointness axioms. Manzano-Macho et al. [6] focus on small corpora of high quality domain text, our system however uses noisy and evolving data from the Web and also includes more diverse sources such as APIs from social media Websites and a linked data source (DBpedia). In terms of evaluation, we employ user-based evaluation with domain experts (see below), whereas Manzano-Macho et al. [6] compare their results against a reference ontology. Gacitua and Sawyer [8] present a quantitative comparison of technique combinations for concept extraction. Although the goal is similar to our work, they investigate which process pipeline of NLP techniques is most helpful for term extraction from a domain corpus, whereas we study the balancing of term lists stemming from heterogeneous evidence sources.

As mentioned, the skillful combination and balancing of evidence sources is a crucial factor to leverage the potential of heterogeneous sources. Spreading activation, which is a method for searching semantic networks and neural networks, is the key tool to integrate evidence sources in our framework. Spreading activation is also frequently used in information retrieval. In his survey Crestani [5] concludes that spreading activation is capable of providing good results in associative information retrieval.

3 The Ontology Learning Framework

As each ontology is generated from scratch, it is straightforward to measure and compare results obtained by using different settings (regarding the evidence from heterogeneous input sources). The experiments discussed in this paper were conducted with an OL system first published by Liu et al. [11]. The generated ontologies are lightweight [18], most OL systems aim at learning ontologies which make little or no use of axioms (lightweight ontologies) [19].

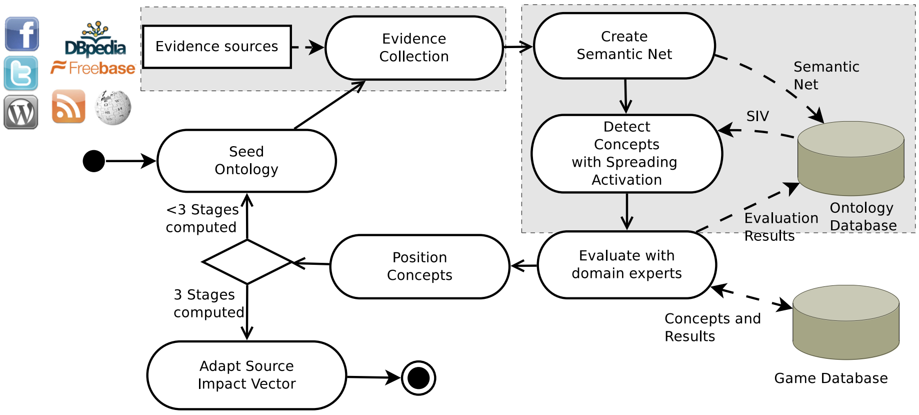

Fig. 1. The Ontology Learning Framework.

As the basic OL framework has been presented before, its description will be kept to a minimum. This section focuses on the new components and elements necessary to understand the evidence integration processes.

The basic workflow of the system, shown in Figure 1, is as follows:

1. The OL starts from a small *seed ontology* – typically only a few concepts, for example `[global_warming subClassOf climate_change.]` in the domain of climate change, in the tennis domain it is `[tennis_match subClassOf tennis.]`.
2. *Collection of evidence* for all seed concepts from the evidence sources (details on the sources of evidence are found in the next section).
3. Integrate the evidence in a so-called *semantic network*.
4. Transform the semantic network into a spreading activation network.
5. The *spreading activation* algorithm yields new concept candidates (concept detection phase).
6. Domain experts *rate* concept candidates as either relevant to the domain or non-relevant.

7. Relation detection and *positioning* of new (relevant) concepts in the ontology.
8. Start over with step one, using the extended ontology as seed ontology in the next iteration. Thereby the ontology gets bigger and more granular.
9. Finally, after a predefined number of extension iterations: Stop.

The parts of Figure 1 highlighted light-gray are the most interesting regarding the evaluation of the system. These parts are either covered in more detail in the upcoming section (evidence sources), or in the remainder of this section.

The neural network technique of spreading activation is a crucial algorithm in the system, used for the selection of new candidate concepts and also in concept positioning. Spreading activation typically activates a number of seed nodes, the algorithm then propagates the activation energy through the network according to link weights. In the iterative process a decay factor D is used to diminish activation propagation farther away from the seed nodes. In concept selection the system simply picks the n candidates with the highest activation level after the spreading activation process has finished, we typically use 25 for n. All evidence is collected with automated methods, which provide some relevant, but also many irrelevant, terms. Irrelevant terms may be hardly or not domain-relevant at all, or too specific, i.e. on a too detailed level of granularity. Spreading activation helps to distinguish relevant terms by integrating all collected information.

The only point in the OL cycle where human intervention is needed is relevance assessment of new concept candidates. It is still unclear if fully automated OL is feasible at all [19]. In the experiments, domain experts evaluated concept candidates. To increase scalability, a component that distributes evaluation tasks to online labor markets (esp. CrowdFlower[1]) is under development.

4 Generation of Evidence

To understand the characteristics of evidence sources, it is necessary to understand the data sources we use, and the methods to extract evidence from these data sources. The evidence sources (listed in Section 4.3) emerge from the application of extraction methods to data sources.

4.1 Heterogeneous Data Sources

In this paper we distinguish between evidence and data sources. Data sources refer to the *raw* resources, they include domain text from various origins, structured data (WordNet, DBpedia), and calls to social media APIs.

Figure 2 provides an overview of the data sources used. Most importantly, the data sources include (i) **Domain text corpora**. Using the webLyzard suite of Web mining tools[2] the framework generates corpora from news media (segregated by geo-location), social media (public postings on Facebook, Youtube, Twitter, etc.) and other sites such as NGO's Websites and the Fortune 1000.

[1] crowdflower.com

[2] www.weblyzard.com

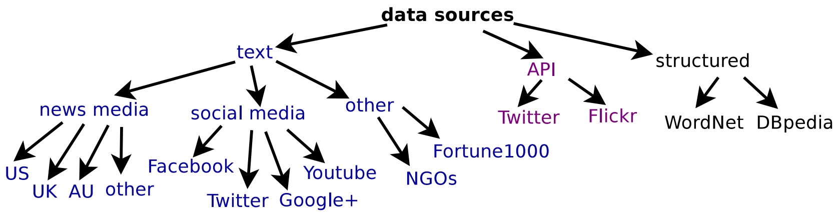

Fig. 2. Heterogeneous data sources used.

Web content typically needs content extraction (boilerplate removal), we apply the approach discussed in [10]. A domain detection tool yields domain-specific documents in the given time interval (month). Domain-detection is only applied for the *climate change* domain, in the *tennis* domain the system uses general (news) media corpora. Furthermore, the system uses **structured sources**, that is WordNet and the DBpedia dataset. And finally, we execute (iii) **API calls** to (social) Web services.

4.2 Extraction Methods

The OL system applies a variety of methods to extract terminology from the data sources – depending on their type. For all text-based sources, we currently use: (i) Computation of keywords for a seed concept (represented by its label). The keyword service (see [11] for details) detects significant phrases and applies co-occurrence statistics to generate a list of keywords ordered by χ^2 significance. The keywords appear in the same *page (document)* or *sentence* as the concept. For short documents (tweets, Facebook postings) we only compute page-level keywords. (ii) Hearst patterns [9], which are lexical patterns to find common phrases that link hypo-/hypernym pairs.

Table 1 includes example data for term extraction, it presents a short snippet of page-level keywords generated for the seed concept "CO2" from UK news media text (evidence source no. 4) in July 2013. This demonstrates the typical characteristics of evidence acquisition: some terms are relevant to the *climate change* domain, some are not relevant, some are too specific. A full listing of evidence for a seed term ("CO2") and examples of ontology run results is found at `https://ai.wu.ac.at/~wohlg/conf_data`. A demo portal of the underlying OL system is available at `http://hugo.ai.wu.ac.at:5050`.

Social Web APIs (Twitter, Flickr) which directly provide related terms (or "tags") are simply queried with a seed concept label as input to extract terminology. These APIs typically provide very recent terminology and are helpful to collect terms complementary to text sources. Example data for social sources is found in Weichselbraun et al. [16].

Finally, regarding structured sources, from WordNet [7] the system extracts hyponyms, hypernyms, and synonyms for an input term, see Liu et al. [11] for

Term	Significance	Term	Significance
carbon price floor	164.85	emission	110.48
sec	135.54	air	99.99
fertilisation	133.63	waste	90.17
PM10	123.45	0-62mph	89.12
environment committee	121.27	flame	86.74
member state	114.62	carbon tax	78.53

Table 1. Example evidence (keywords and their χ^2 co-occurrence significance) for the seed concept "CO2".

details and examples. We also query the DBpedia SPARQL endpoint[3] with *dbpedia:acronyms*, *dcterms:subject* and *dbpedia:othernames* predicates to get related terms to a seed term. For our example of term "CO2", *dcterms:subject* suggests the following terms from DBpedia: "Acid anhydrides", "Acidic oxides", "Carbon dioxide", "Coolants", "Fire suppression agents", "Greenhouse gases", etc.

4.3 Evidence Sources

This section primarily gives an overview about all 32 heterogeneous sources of evidence used in the experiments with an OL system. As already mentioned, evidence sources arise from the application of extraction methods on data sources. Table 2 presents the 26 evidence sources originating from using the keyword and Hearst pattern techniques on domain text data sources.

	Method		
Data sources			
domain text from:	Keywords/page	Keywords/sentence	Hearst patterns
US news media	1	2	3
UK news media	4	5	6
AU/NZ news media	7	8	9
other news media	10	11	12
Social media: Twitter	13	-	14
Social media: Youtube	15	-	16
Social media: Facebook	17	-	18
Social media: Google+	19	-	20
NGOs Websites	21	22	23
Fortune 1000 Websites	24	25	26

Table 2. The 26 evidences sources used in the ontology learning process based on domain text. The data is collected from the Web to create corpora in monthly intervals.

Every line in Table 2 represents a data source. Each text data source (except the ones with very short documents), yields three evidence sources, namely

[3] dbpedia.org/sparql

keywords on page (document) level, keywords on sentence level, and relations (and terms) extracted with Hearst patterns [9]. 26 of 32 evidence sources extract terminology from text, making domain text corpora an important input to the system.

	Method				
Data source:	hypernyms	hyponyms	synonyms	API	SPARQL
WordNet	27	28	29	-	-
DBpedia	-	-	-	-	30
Twitter	-	-	-	31	-
Flickr	-	-	-	32	-

Table 3. The remaining 6 evidence sources, which are based on WordNet, Social Media APIs, and DBpedia.

The remaining 6 evidence sources extract terms from WordNet and DBpedia, or with social Web API queries as shown in Table 3.

Obviously, the 32 sources are heterogeneous in type and number of results, we use spreading activation to integrate evidence (see Section 3) and parameters to balance and limit the number of evidences per source (see below).

5 Evaluation

This section includes evaluation results of ontology learning (OL) experiments conducted between July 2013 and December 2014. Starting from the seed ontology the system generated 75 concept candidates (3 runs of 25 concepts each) per ontology – this fixed number of 75 concept candidates per ontology was used in all upcoming experiments, irrespective of the number of evidence sources used. After Section 5.1 provides details about the evidence (term lists) used, Section 5.2 describes the experiments for integrating heterogeneous evidence. Section 5.3 discusses concept relevance assessment.

5.1 Characteristics of Evidence Sources

In order to get a meaningful interpretation of evidence balancing and integration, first the characteristics of the underlying input data need to be investigated. Two properties greatly vary between evidence sources: the number of evidences provided (for a seed concept), and the average term quality per evidence acquisition method. Term quality was measured as the ratio of terms supplied by the respectively method which label a relevant domain concept. A domain expert manually evaluated sufficiently large term lists for different seed concepts and methods – resulting in a few thousand terms – to assess term quality.

Table 4 gives an overview of these characteristics. It lists the methods described in Section 4, and gives the rough average numbers of evidences per seed

		Term Quality		
Method:	Avg. Num. of Evid.	Top 25	Top 100	Top 500
Keywords/page	400	0.31	0.26	0.12
Keywords/sentence	200	0.27	0.19	0.10
Hearst Patterns	18	0.15		
API Twitter	70	0.10		
API Flickr	16	0.18		
WordNet (Hypernyms)	15	0.24		
WordNet (Hyponyms)	17	0.21		
DBpedia	13	0.27		

Table 4. Average number of evidence and evidence quality per extraction method.

concept which the evidence sources provide (*Avg. Num. of Evid.*). Furthermore, the table includes term quality in the remaining columns. Only co-occurrence statistics-based terms (*keywords*) have a significance value assigned (and are thereby ordered), for these we evaluated the top 25, top 100, and top 500 most significant terms. For all other sources we evaluated all terms supplied. Table 4 shows (i) that the average number of evidences greatly differs between sources, and also that term quality varies to a large extent. Term quality is high for the 25 most significant keywords per seed concept, and also for terms provided by WordNet and DBpedia. Keywords of low significance, and social sources (esp. Twitter) yield low quality terms on average. Hearst patterns generate rather sparse results which are of moderate quality.

One aspect of using heterogeneous sources is that they provide **complementary input** to better cover the domain of interest. In our system, corpus-based techniques (mostly keywords) account for the base layer of evidence. Apart from text-based input, social sources add very recent and emotional terminology, helpful to improve results and capture dynamic aspects of the domain [16], but also include a large share of noise, typos, etc. WordNet typically offers general and high quality input, which also helps to build the taxonomic backbone, but does not reflect dynamic aspects of domain evolution. The current version of SPARQL queries against DBpedia returns specific and technical terms, but also many terms which are too specific or not relevant to the domain.

Balancing the number of evidences. As seen in Table 4, if not limited, the number of evidences (terms) supplied strongly varies between evidence sources. Whereas the number of keywords for a concept sometimes exceeds 1000 terms, other sources provide comparably few results. In the upcoming section we present experiments where the number of evidences per source is either not limited, or limited to a maximum number of evidences per source to (i) balance the influence of sources on the resulting ontology and (ii) study which impact the amount of evidence has on the quality of concept candidates suggested by the OL system.

5.2 Leveraging and Balancing Sources and Evidences

As stated in Section 1, the goal of this research is to provide hints and insights on the combination and integration of heterogeneous evidence sources in OL (specifically for the concept detection phase) which can be generalized.

$$Accuracy = \frac{Relevant\ concept\ candidates\ generated}{All\ concept\ candidates\ generated} \tag{1}$$

In this section, we measured the accuracy of the system by the ratio of relevant concept candidates resulting from the OL system, see Equation 1. Other aspects, such as the positioning of new concepts in the ontology and the detection and labeling of relations are not part of this study, some of these points are covered in [17].

The Number of Evidences per Source. The first question to address is the impact of the number of evidences per source on the quality of concept candidates. Table 5 summarizes experiments where every of the 32 evidence sources was limited to suggest only 5, 10, etc. evidences per seed concept. As discussed in the previous section, some sources like WordNet or DBpedia typically provide very few evidence, whereas keyword-based sources produce up to 1000 terms per source. Obviously, limiting all sources to (for example) 10 evidences per seed concept, will reduce the impact of keyword-based sources. Using limits i) balances to number of evidence between sources, ii) saves computation time, but also iii) removes data which might be helpful in the spreading activation (ie. evidence integration) process. We use two domains in the experiments, *climate change* and *tennis*. The *climate change* ontologies were generated from scratch in every month between July 2013 and November 2014, the *tennis* ontologies between July 2014 and November 2014. The accuracy numbers in Table 5 are based on 17 ontologies computed per respective setting for *climate change*, which leads to 1275 ($75 * 17$) concept candidates per setting. In the *tennis* domain, we have 5 ontologies per setting with 375 concept candidates. If not stated otherwise, these numbers also apply to upcoming tables later in this section.

With very few evidences per source (*limit=5*), the benefits of redundancy and integration of heterogeneous sources are small (poor accuracy), although using only the best (most significant) keywords. Only in interactive systems where runtime is a very critical issue such a setting should be considered. On the other hand, in our experiments with a limit of 200 or more evidences per seed concept, the number of evidences per source is unbalanced, and more and more keywords with low significance are added to the spreading algorithm network, negative effects exceed the benefits of additional evidence data. Accuracy is lower in the *tennis* domain, we attribute this to the underlying data used, which are general domain-agnostic (news) media corpora, whereas for *climate change* the system uses domain-specific corpora.

In contrast to our initial expectations that more evidence is always better regarding resulting ontology quality (although it will be computationally expensive), even low numbers (*limit=10*) allow high accuracy if evidence is ranked

No. of Evidences	Acc. CC	Acc. Tennis	Acc. Random Keyw. CC
limit=5	56.44	46.80	52.72
limit=10	64.05	55.53	56.51
limit=20	67.57	60.27	60.98
limit=50	68.68	59.87	61.64
limit=100	67.79	58.27	62.73
limit=200	67.87	58.53	65.13
limit=500	66.39	57.88	66.01
no limit	66.29	57.34	66.29

Table 5. Accuracy of concept detection (percentage of relevant concept candidates) for the domains of *climate change* (CC) and *tennis* depending on the number of evidences per source, with default and random selection of keyword evidence.

by expected quality. In our system keywords are ranked by their significance value. Our experiments suggest that in the range of 20 to 50 terms per evidence source very good or even best results can be expected. However, this is only true while using a sufficient number of evidence sources (see below). A remark: the differences in accuracy in Table 5 are statistically significant, eg. with $p = 0.009$ between accuracy of *limit=10* and *limit=20*.

A more detailed look at the ontologies exhibits a more frequent occurrence of specific and exotic (but still relevant) concepts when using a low limit (such as *limit=5*), while a high limit promotes more general terms. This fact, which is in favor of high *limit* settings is not reflected by the data in Table 5.

Out of curiosity we also experimented with choosing keywords randomly from the list of all keywords (instead using of the most significant), see column *Acc. Random Keyw. CC* in Table 5. As expected this lowers the accuracy with low *limits*, and gives a more realistic picture for systems where evidence per source is not ordered. Therefore, in a machine learning environment where the expected quality of evidence is unknown and there is no explicit grading, it is advisable to use more evidence per source to fully benefit from redundancy. Another experiment, in which the keyword significance (as yielded by co-occurrence statistics) was not used at all, gave very poor results. This confirms that the quality of sources is important, and that low-quality evidence cannot be compensated by using multiple sources entirely.

In summary, it is important to have enough data to benefit from redundancy and aggregation. Additional evidence beyond this point can even have a negative impact if the balance between sources is lost, or the quality of additional evidence is not sufficient.

The Number of Evidence Sources Used. Not only the number of evidences per source is important, also the influence of the number of heterogeneous sources on the learning algorithm has to be taken into consideration. We evaluated the impact of using (i) only one source which yields rather low quality terms (*1 Twitter*), (ii) only one source with high quality input (page-level keywords from UK media – *1 UK-KW-page*), (iii) five random sources (*5 sources*), (iv) *15 sources*, (v) all sources (*32 sources*). Table 6 presents the results for these five variants, it shows the outcome for limit settings with the number of evidences (terms) not exceeding 50 and 200, respectively.

%Relevant	**1 (Twitter)**	**1 (UK-KW-page)**	**5 srcs**	**15 srcs**	**32 srcs**
CC limit=50	16.54	48.80	59.52	68.28	68.84
CC limit=200	19.85	49.78	57.48	67.73	67.64
Tennis limit=50	21.15	50.67	52.25	56.88	57.87
Tennis limit=200	23.17	52.78	54.33	57.74	58.33

Table 6. Accuracy (percentage of relevant concept candidates) of concept detection regarding the number of evidence sources ("srcs") used – for two limit-settings, in the domains of *climate change* and *tennis*.

When relying on a single source, the quality of evidence of that source is essential, obviously – see *1 Twitter* and *1 UK-KW-Page*. In our experiments, *5 sources* of mixed quality are sufficient to see the benefits of using multiple sources. Around *15 sources* can be enough to gain the full advantage of heterogeneous evidence integration and redundancy. This means that a small and computationally efficient spreading activation network with a sum of a few thousand terms (suggested by 10-15 sources) can be quite enough to get best results. The difference between 5 and 15 evidence sources is statistically significant e.g. for the *climate change* domain as confirmed with a binomial test ($p \approx 0.0006$ for both *limit* settings).

The know-how regarding the minimal number of sources necessary can be helpful in various situations: (i) when setting up a new system, (ii) when there is need to scale down an existing system that is too slow or consumes too many resources, (iii) when there is need to use only a subset of evidence sources for a particular application. For example, we plan ontology evolution and trend detection experiments in which we will only use sources which are highly dynamic, and omit more static sources such as WordNet.

The Number of Seed Concepts. Finally, we investigated the impact of the type and number of seed concepts for which evidence is collected. Our system learns ontologies in 3 iterations of extension. In the first iteration (*Stage1*) there are only very few seed concepts (in the climate domain: "climate change" and "global warming"), which are obviously very relevant to the domain. The seed concepts in *Stage2* are the expert confirmed concepts learned in *Stage1*, in *Stage3* the system uses the concepts acquired in *Stage2*. The concepts in *Stage2* are typically more general than in in *Stage3*, in which the ontology gets more granular. Table 7 presents the ratio of relevant concepts suggested regarding the stage (the number and granularity of seed concepts) and the number of evidences used.

	Stage1 – 2 SC	**Stage2 – ca. 18 SC**	**Stage3 – ca. 35 SC**
limit=5	54.67	61.87	56.53
limit=50	80.30	69.96	55.56
limit=200	78.83	68.33	56.22

Table 7. Accuracy depending on seed concepts (SC) and evidence limit applied.

A combination of a low number of seed concepts in *Stage1* and low number of evidences (*limit=5*) does not provide spreading activation with enough data to produce good results. Such a setting creates a network with only a few hundred evidences (2 SC $*$ 32 sources $*$ 5 evidences per source). This is well below the critical number of evidences of a few thousand (according to our experiments) which is needed for high accuracy. On the other hand, when the number of seed concepts is high, then a high number of evidences per seed concept offers no additional benefit, accuracy for *limit=50* and *limit=200* is very similar. The best results are achieved in *Stage1*, which uses domain concepts of high relevance and generality, and enough evidence to exploit redundancy (limit $\geq$ 50). The accuracy in Stages 2 and 3 is diminishing, because the seed concepts tend to get less domain-relevant with increasing distance from the initial seed ontology.

Observations. A list of key observations and hints concludes this section: (i) It is critical to ensure having enough evidence to benefit from redundancy at every step of the learning cycle. In our system enough evidence corresponds to at least a few thousand pieces of evidence (terms). Additional evidence beyond this points only slows the system down while providing little use. (ii) When there is no order (regarding quality) in evidence data, then more evidence will be needed to get the best results. (iii) Using our evidence integration method and settings, around 10-15 sources of heterogeneous evidence are sufficient to gain the full effect of evidence integration. (iv) Balancing input from evidence sources is typically more important than the raw number of evidence per source.

And interesting strain of future work will be the attempt to optimize the source impact vector (SIV), which controls the influence of a particular source on the learning process. In this research we use a uniform source impact for all evidence sources as the goal is to study the balancing of evidence. In future work we will try to find an (almost) optimal configuration of impact of evidence sources in the spreading activation network. Preliminary studies show that this will lead to a significantly higher accuracy of the system.

5.3 Relevance Assessment

This section, which concludes the evaluation, takes an alternative view at the judgments on concept relevance made by the domain experts, especially on concept candidates rated *non-relevant.*When rating concept candidates, domain experts had only two choices: *relevant* or *non-relevant* to the domain at the given level of granularity. We took a more detailed look at concept candidates that were rated as *non-relevant.* From 100 candidates rated *non-relevant* to the domain of *climate change*, 61% were in fact at least partly relevant to the domain, but very generic or too specific. Only the remaining 39% were not relevant at all, but according to the domain experts not relevant for this level of granularity. For this reason, depending on point of view and granularity, the accuracy of the OL system is higher than stated in the evaluation data. Among the 61% of candidates partly relevant to the domain of *climate change* are mostly terms that are too generic (for example: “impact”, “mitigation”, “issue”, “policy”, etc.). The 39% of clearly non-relevant terms include fragments from the phrase detection algorithm such as “change conference” or candidates simply unrelated (“century”, “level”, “wave”).

6 CONCLUSIONS

The integration of heterogeneous evidence sources can improve accuracy in ontology learning and other areas which use similar machine learning techniques and multiple evidence sources. In this paper we study how a system needs to be set up to gain the desired results, and give hints and insights on the impact on accuracy of the number of evidences per source, the number of evidence sources, of quality per evidence source, etc. Among the key findings and contributions is the surprising fact that a limited number of evidences – a few thousand terms from heterogeneous sources – provides results of similar quality compared to using much higher numbers. In addition, in our experiments around 10-15 evidence sources were sufficient to gain full benefits of redundancy and evidence aggregation. Heterogeneous sources of evidence not only help to raise accuracy, but also offer complementary vocabulary to cover the domain.

Future work will apply the presented experiments to similar systems. We expect similar results as the basic characteristics of evidence integration do not change. Furthermore, we will further optimize the system using the source impact vector (SIV) by (i) adapting the SIV over time according to the quality of concept candidates suggested by the source to increase the impact of sources that consistently suggest a high ratio of relevant concepts, and (ii) conducting optimization experiments of find an optimal configuration for the SIV.

Acknowledgments. The work presented was developed within project uComp, which receives the funding support of EPSRC EP/K017896/1, FWF 1097-N23, and ANR-12-CHRI-0003-03, in the framework of the CHIST-ERA ERA-NET programme.

References

1. Alani, H.: Position paper: ontology construction from online ontologies. In: Carr, L., Roure, D.D., Iyengar, A., Goble, C.A., Dahlin, M. (eds.) WWW 2006, Edinburgh, Scotland, May 23-26. pp. 491–495. ACM (2006)

2. Buitelaar, P., Cimiano, P., Magnini, B.: Ontology Learning from Text: An Overview, vol. 123, chap. 1, pp. 3–12. IOS Press (7 2005)
3. Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S.: Ontology Learning from Text, chap. Learning Taxonomic Relations from Heterogeneous Sources of Evidence, pp. 59–76. IOS Press, Amsterdam (2005)
4. Cimiano, P., Völker, J.: Text2onto. In: Montoyo, A., Muñoz, R., Métais, E. (eds.) NLDB. Lecture Notes in Computer Science, vol. 3513, pp. 227–238. Springer (2005)
5. Crestani, F.: Application of spreading activation techniques in information retrieval. Artificial Intelligence Review 11(6), 453–482 (1997)
6. David Manzano-Macho, A.G.P., Borrajo, D.: Unsupervised and domain independent ontology learning: Combining heterogeneous sources of evidence. In: Nicoletta Calzolari, Khalid Choukri, e.a. (ed.) Proceedings of LREC'08. European Language Resources Association (ELRA), Marrakech, Morocco (May 2008)
7. Fellbaum, C.: Wordnet an electronic lexical database. Computational Linguistics 25(2), 292–296 (1998)
8. Gacitua, R., Sawyer, P.: Ensemble methods for ontology learning - an empirical experiment to evaluate combinations of concept acquisition techniques. In: ICIS 08. pp. 328–333. IEEE Publishing (5 2008)
9. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: COLING'92. pp. 539–545. Nantes, France (1992)
10. Lang, H.P., Wohlgenannt, G., Weichselbraun, A.: Textsweeper - a system for content extraction and overview page detection. In: Int. Conference on Information Resources Management (Conf-IRM). pp. 17–22. AIS, Vienna, Austria (2012)
11. Liu, W., Weichselbraun, A., Scharl, A., Chang, E.: Semi-automatic ontology extension using spreading activation. JUKM 0(1), 50–58 (2005)
12. Sanchez, D., Moreno, A.: Learning Non-taxonomic Relationships from Web Documents for a Domain Ontology Construction. DKE 64(3), 600–623 (2008)
13. Tatarintseva, O., Ermolayev, V., Keller, B., Matzke, W.E.: Quantifying ontology fitness in ontoelect using saturation- and vote-based metrics. In: Ermolayev, V., Mayr, H., Nikitchenko, M., Spivakovsky, A., Zholtkevych, G. (eds.) Information and Communication Technologies in Education, Research, and Industrial Applications, CCIS, vol. 412, pp. 136–162. Springer (2013)
14. Velardi, P., Faralli, S., Navigli, R.: OntoLearn Reloaded: A Graph-based Algorithm for Taxonomy Induction. Computational Linguistics 39(3), 665–707 (2013)
15. Völker, J., Vrandečić, D., Sure, Y., Hotho, A.: Learning disjointness. In: Proceedings of the 4th European Conference on The Semantic Web: Research and Applications. pp. 175–189. ESWC '07, Springer-Verlag, Berlin, Heidelberg (2007)
16. Weichselbraun, A., Wohlgenannt, G., Scharl, A.: Augmenting lightweight domain ontologies with social evidence sources. In: Tjoa, A.M., Wagner, R.R. (eds.) DEXA 2010. pp. 193–197. IEEE, Bilbao, Spain (August 2010)
17. Weichselbraun, A., Wohlgenannt, G., Scharl, A.: Refining non-taxonomic relation labels with external structured data to support ontology learning. Data & Knowledge Engineering 69(8), 763–778 (2010)
18. Wohlgenannt, G., Weichselbraun, A., Scharl, A., Sabou, M.: Confidence management for learning ontologies from dynamic web sources. In: KEOD 2012. pp. 172–177. SciTePress, Barcelona, Spain (October 2012)
19. Wong, W., Liu, W., Bennamoun, M.: Ontology learning from text: A look back and into the future. ACM Computing Surveys 44(4), 20:1–20:36 (Sep 2012)