Ontology-Based Faceted Semantic Search with Automatic Sense Disambiguation for Bioenergy Domain

Feroz Farazi†, Craig Chapman†, Pathmeswaran Raju† and Lynsey Melville±

†Knowledge Based Engineering (KBE) Lab
±Centre for Low Carbon Research
Faculty of Computing, Engineering and the Built Environment
Birmingham City University
City Centre Campus, Millennium Point, Birmingham, B4 7XG, UK
E-mail: {mohammad.farazi, craig.chapman, path.raju}@bcu.ac.uk
E-mail: lynsey.melville@bcu.ac.uk

Abstract: WordNet is a lexicon widely known and used as an ontological resource hosting comparatively large collection of semantically interconnected words. Use of such resources produces meaningful results and improves users' search experience through the increased precision and recall. This paper presents our facet-enabled WordNet powered semantic search work done in the context of the bioenergy domain. The main hurdle to achieving the expected result was sense disambiguation further complicated by the occasional fine-grained distinction of meanings of the terms in WordNet. To overcome this issue, this paper proposes a sense disambiguation methodology that uses bioenergy domain related ontologies (extracted from WordNet automatically), WordNet concept hierarchy and term sense rank.

Keywords: semantic search; faceted search; faceted semantic search; Knowledg Base; WordNet; ontology; bioenergy.

Biographical notes: Dr Feroz Farazi is a Research Fellow performing full-time research in KBE, ontology and semantics for engineering and robotics. He was working as a full-time researcher in the EnAlgae project funded by the INTERREG IVB North West Europe programme. In this project he was contributing in the area of Knowledge Based Decision Support Systems. He is actively participating in the research activities that deal with data semantics and ontology based tools. The research activity of Dr Farazi also includes effective use of referential semantics in developing pilot scale and lab scale decision support tools and technologies and quantification of the potential added value through the inclusion of semantics and ontology in design automation.

Professor Craig Chapman is Senior Academic for Research in the School of Engineering, Design and Manufacturing Systems. He has worked at an international level working in Europe, USA and the UK, holding positions in industry from Director, Principal Design Engineer, Design Group Leader and Senior Applications consultant. In Academia Craig's career has taken him from Senior Research Fellow, Head of the Knowledge Based Product Development Lab at Warwick Manufacturing Group, University of Warwick to Head of the Knowledge Based Engineering Lab and Senior Academic for Research at Birmingham City University. The main focus of Craig's career has been research, development and the application of design engineering automation and the development of Knowledge Based Engineering solutions, enabling companies to rapidly respond to design engineering changes and explore multiple solutions with consideration to all life cycle phases.

Dr Pathmeswaran Raju is a Reader in Knowledge Modelling and Engineering at Birmingham City University. His current research interests include knowledge engineering for design, engineering and manufacturing disciplines for the development of knowledge-based engineering systems. He was involved with developing knowledge based decision support system for the Energetic Algae (EnAlgae) INTERREG IVB NWE Programme, and developing requirements and knowledge models for Platform Independent Knowledge Model (PIKM) project as part of Technology Strategy Board (TSB)/Rolls-Royce funded Strategic Investment in Low-carbon Engine Technology 2 (SILOET2) programme and EU Clean Sky initiative.

Dr Lynsey Melville is a Reader in Bioenergy and Director of the Centre for Low Carbon Research. She is currently leading the Bioenergy research group whose focus is on accelerating the adoption of environmentally sustainable and commercially viable energy from biomass. This involves supervision of both post doctoral and PhD researchers. Lynsey was principle investigator on two large EU funded programmes – EnAlgae and BioenNW. Working closely with the knowledge based engineering group and a broad range of stakeholders. The aim of this work was to capture data and information from across the whole bioenergy delivery chain. One of the outcomes of this work was a number of innovative and adaptive decision support systems, which enable stakeholders to identify optimal sites, partners and markets and develop project plans which can be tailored to regional conditions.

1 Introduction

Search engines have been in place for decades and search giants including Google (https://www.google.com/), Bing (https://www.bing.com/) and Yahoo! (https://www.yahoo.com/) are playing a significant role in fulfilling the information needs of users. Many users would not want to imagine how the World Wide Web would be without their presence. Continuous improvement efforts made for behind these tools make them offering consistently better search results.

The syntactic search approach, which takes into account the presence of the query term(s) in the target set of documents and returns only those ones that contain one or more query terms (Lei et al., 2006), has been used from the beginning of the search engines era. This approach is too rigid and narrow, leads to documents containing *algae cultivation for ethanol production* not being picked up when the search term is *algae cultivation for ethyl alcohol production* or *algae cultivation for fermentation alcohol production* (even though *ethanol*, *ethyl alcohol* and *fermentation alcohol* are the alternative names of the same agent). In WordNet (https://wordnet.princeton.edu/), *Ethanol* is defined as follows.

> *{ethyl alcohol, ethanol, fermentation alcohol} -- (the intoxicating agent used pure or denatured as a solvent; proposed as a renewable clean-burning additive to gasoline)*

In this example, the terms which share the same meaning are enclosed in braces ({}) and separated by commas. The textual description which conveys human readable semantics is given in parentheses. Another similar example of sharing the same meaning by multiple terms is provided below.

> *{United Kingdom, UK, U.K., Great Britain, GB, Britain} -- (a monarchy in northwestern Europe; divided into England and Scotland and Wales and Northern Ireland)*

However, there is a difference between the former example and the latter. The former represents a concept or class and the latter, on the other hand, represents an instance or entity. Each term in the former case is an alternative name of the same concept and each term in the latter is an alternative name of the same entity. Alternative names of the same concept can be represented with the 'semantically equivalent' or *owl:equivalentClass* (Dean and Schreiber, 2004) relation and alternative names of the same entity can be represented with the 'same as' or *owl:sameAs* (Dean and Schreiber, 2004) relation.

The semantic search approach nullifies the need for explicit appearance of a search term in the documents. Instead, the presence of the concept or

entity that is semantically equivalent to the concept or entity of the search term is sufficient. Semantic search can check much deeper and produce insightful results by taking into account all possible relations of the concept or entity users are interested in (Guha et al., 2003; Lei et al. 2006; Giunchiglia et a., 2009; Zhong et al., 2002). As the semantic search approach goes beyond the capabilities of what syntactic search can offer us now, it is employed in a number of different applications including Web service retrieval (Srinivasan et al., 2004), multimedia object identification (Schreiber et al., 2008; Chang et al., 2007), bio-medical data retrieval (Kaniovskyi et al., 2015) and wiki page search (Hasse et al. 2008; Schaffert, 2006; Auer et al., 2006). There are attempts to apply this approach in settings different from the Web, e.g., desktop search (Chirita, 2005).

Often ontologies are put in place to enable the semantic capabilities of applications resulting semantic search (Hyvonen, 2004; Bonino, 2004; Fang et al., 2005). Ontological resources are built with semantic relations using more generic, more specific and equivalent relations.

WordNet is an ontology that showed its potential in the implementation of semantic search in both domain dependent applications (Buscaldi et al., 2005) and domain independent ones (Kruse et al., 2005). It consists of semantic relations such as *synonymy* or *equivalent*, *hypernymy* or *is-a* and *holonymy* or *part-of* (Miller, 1995). This work focuses on the use of (the semantically) *equivalent* relation for the performance boost, in terms of both precision and recall, in our already developed domain dependent faceted syntactic search application.

Our work is situated in the context of the EnAlgae[1] project, an Interregional North West Europe funded project within the bioenergy domain and investigating alternative renewable green energy producing initiatives. EnAlgae is an acronym for Energetic Algae. This project investigates the feasibility of gaining energy (e.g., *methane*, *ethanol* and *biodiesel*) from different kinds of *algae* and makes analytical data, reports and economic models (Sapkota et al., 2015) available in the form of documents for stakeholders. We have extended the syntactic search application with semantic capability to provide more relevant documents to the stakeholders regardless of the variations in wording forming the query.

To enable semantics in the search, the ontological constituents of WordNet are used. The semantic search application should behave the same for semantically equivalent terms and should identify the same set of

[1] http://www.enalgae.eu/

documents when a term in a query is replaced by its semantically equivalent counterpart. Therefore, the queries *algae cultivation in UK* and *algae cultivation in Britain* should return the same result. *UK* is present in the list of country/region metadata extracted for developing the syntactic search application but *Britain* is not. Given this, the semantic search application can retrieve the result. It should retrieve the same set of documents for the following queries as well: *algae cultivation in United Kingdom*, *algae cultivation in Great Britain* and *algae cultivation in GB*. Our assumption remains true for all queries except in the case of the last one where *GB* is present.

The reason for this deviation is due to the *polysemy* in WordNet. Polysemy refers to the multiple meanings (senses) of a term. Taking the sense with the highest rank gives better accuracy (Suchanek et al., 2007). Note that in WordNet senses are explicitly ranked. *United Kingdom*, *UK* and *Great Britain* are there with their country sense ranked highest, *GB* has four senses where the country sense ranked lower. As a result, it was not picked up. *UK* was not taken as one of the equivalent terms for performing the search. Rather its *sarin* sense was taken as the synonym, which is not the intended one in this case. For the purpose of clarity, the senses of GB are reported as follows:

{sarin, GB} -- (a highly toxic chemical nerve agent that inhibits the activity of cholinesterase)

{gilbert, Gb, Gi} -- (a unit of magnetomotive force equal to 0.7958 ampere-turns)

{gigabyte, G, GB} -- (a unit of information equal to one billion (1,073,741,824) bytes or 1024 megabytes)

Out of four, here, we reported three senses of GB. The other represents its country sense (Great Britain) described earlier in this Section. Note that the senses are included here in accordance with their rank. The sense appearing first in the list has the highest rank, the one appearing second has the second highest rank, and so on. The country sense of GB would appear as the last in this listing as it has the lowest rank. Though the highest ranked sense selection approach offered us correct result in the case of *Britain* and *Great Britain*, the sense disambiguation issue still remains as it failed dealing with *GB*.

To cope with this situation and give better experience to users, we have developed a sense disambiguation tool that uses WordNet semantic relations, domain ontologies (such as plant, chemistry and finance) and term sense rank. The novelty of our approach is that it can automatically identify and extract the domain ontologies needed for our application from WordNet. Therefore, this approach can be replicated in settings where the

applications might need different sets of domain ontologies. In addition to this, our approach supports automatic update of the ontologies. New versions of WordNet, if backward compatible, can be accommodated with no additional development cost. This paper makes the following contributions:

i) The development of an algorithm that can determine and extract domain ontologies from WordNet and order them in terms of relevancy.

ii) The development of a sense disambiguation methodology that takes into account the WordNet synset hierarchies built with semantic relations, relevancy of the domain ontologies and term sense rank.

iii) The creation of a semantic search application that can help fulfill the information needs of the stakeholders in the bioenergy domain.

Section 2 provides a brief description of WordNet with an emphasis on knowledge organisation and domain relations. Section 3 details the domain ontology identification and extraction procedure. Section 4 shows how the sense disambiguation is performed and Section 5 demonstrates the semantic search application. Section 6 reports on the experimental results and evaluation and Section 7 covers the related work. In Section 8, the paper concludes with some possibilities for the future works.

2 WordNet

WordNet is a manually built large lexical Knowledge Base (KB) developed at Princeton under the direction of George A. Miller (Miller, 1995). From this point onward in this paper, WordNet and WordNet KB are used alternatively. In the following subsections, the knowledge organisations and domain relations of WordNet are briefly described.

2.1 Knowledge Organisation

WordNet consists of words, synsets and relations. Each word has a meaning and words with the same meaning are grouped together and called a synset. A synset can also be defined as a set of synonymous words. For example, United Kingdom and Britain are synonymous and they belong to the same synset.

Synsets with more specific meanings are put under the ones with more generic meanings. Note that due to the limited space all the words of a synset are not listed in the following examples. The relation between a

more specific synset and a more generic one is called hypernymy, for example, *{United Kingdom}* has hypernymy *{country}* and the relation between a more generic synset and a more specific one is called hyponymy, for example, *{country}* has hyponymy *{United Kingdom}*. Though hypernymy is an inverse relation of hyponymy, they are explicitly codified.

Synsets can even be part of some other synsets. The relation between two synsets, one treated as a part and another treated as a whole, called holonymy, for example, *{United Kingdom}* has holonymy *{European Union}*. The relation between two synsets, one treated as a whole and another treated as a part, is called meronymy, for example, *{European Union}* has meronymy *{United Kingdom}*. Similar to hypernymy and hyponymy, holonymy is an inverse relation of meronymy and they are also made explicit.

2.2 Domain Relations

In WordNet, domains are defined explicitly with specific kinds of relations linking a synset representing a domain to the member synsets and vice versa. There are three relations and their inverses forming a set of six relations constructing the domain networks. The domain networks are of type topic (e.g., chemistry and finance), region (e.g., United Kingdom and Belgium) and usage (e.g., trade name and idiom). Each network has two relations: domain of synset and member of this domain. One is the inverse of the other.

3 Domain Ontology

Domain ontology is an ontology capturing knowledge about a topic or region or usage. Domain ontologies of WordNet are not well balanced though they can be used in natural language processing tasks (Bentivogli et al., 2004) and making semantically interoperable systems (Giunchiglia et al., 2010; Ganbold et al., 2014; Maltese and Farazi, 2011). In this paper, domain ontology is also referred to as domain. WordNet synsets are grouped into four grammatical categories: noun, verb, adjective and adverb. This work deals with noun category in which 520 domains are available. As all noun synsets in WordNet are hierarchically organized with *is-a* subsumption relation and occasionally with *part-of*. We have exploited this feature and taken for granted that all the synsets which are more specific than the synset of a domain are the members of that domain. Therefore, in the domains there are members coming from the concept hierarchy and original WordNet domain networks.

Each figure from 1(a) to 1(c) depicts a subset of one of the three major domain categories. In these figures, domains are accompanied by the statistics of the noun terms belonging to them.

In Figure 1(a), a subset of 424 topic domains including organism, plant and vegetation are shown. The organism domain consists of 43,850 terms. The size of the plant domain is less than half of the organism domain. The vegetation domain is a quarter of the size of the plant domain.

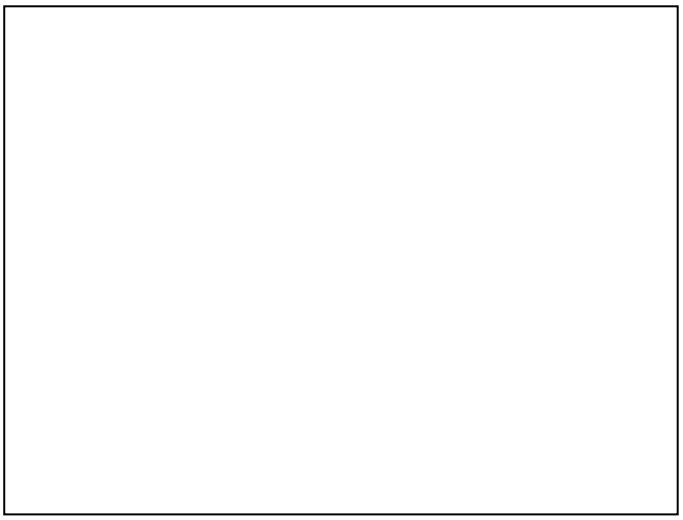

Figure 1(a) Topic domain ontologies (a subset)

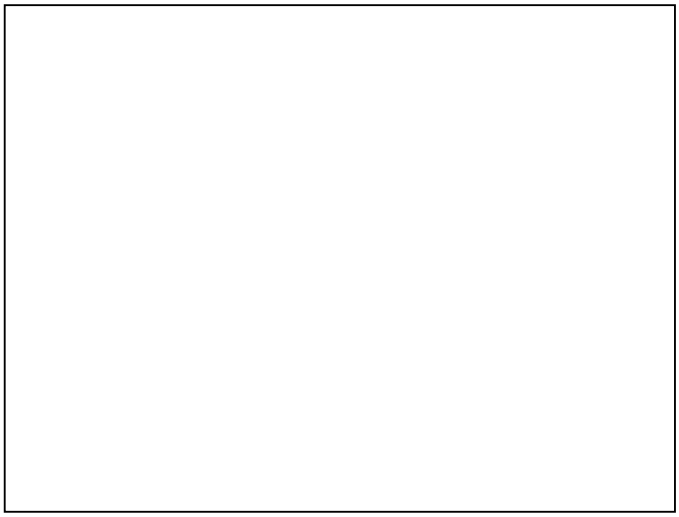

Figure 1(b) Region domain ontologies (a subset)

Figure 1(b) presents a subset of 16 region domains, where Europe,

United Kingdom and France consist of 2676, 496 and 294 terms, respectively.

Figure 1(c) shows that wit, disparagement and trope are domains under the usage category and they have 182, 70 and 50 terms, respectively.

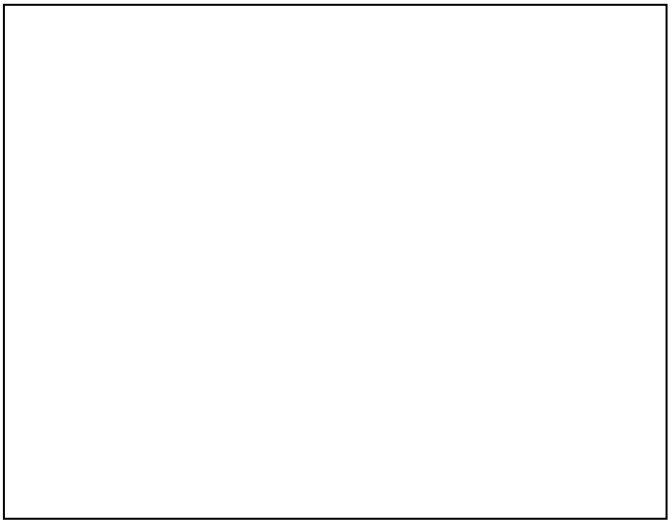

Figure 1(c) Usage domain ontologies (a subset)

Though the size of the domains varies a lot, a closer look at the terms reveals the fact that the terms are organised and clustered meaningfully together. For example, plant domain contains terms such as crop, aquatic and acrogen; vegetation domain contains terms such as bush, grove and shrubbery; and chemistry domain contains co2, ethanol and protein.

4 Semantic Enrichment

There are terms which are interchangeable and leave the meaning of a sentence unchanged. These are called semantically equivalent terms. For example, Netherlands and Holland are semantically equivalent terms. This section, describes how the semantically equivalent term(s) are computed. As shown in Figure 2 all noun domains are extracted from WordNet KB. All terms except the stop words (e.g., a, an, in, etc.) are extracted from the documents on which the search is performed.

The relevant domains are identified by checking for the presence of each extracted term in the whole set of domains. Three possible situations can arise. It can be the case that a term is present in a domain or it is present in multiple domains or it is not present at all in any of the domains. It maintains a term vs domain(s) matrix. In this matrix, it also puts the terms not present in any of the domains. Finally, it identifies the domains,

which are present in the matrix and in turn become relevant for the semantic enrichment. Between the two domains, the one that contains the higher number of terms is more relevant than the other.

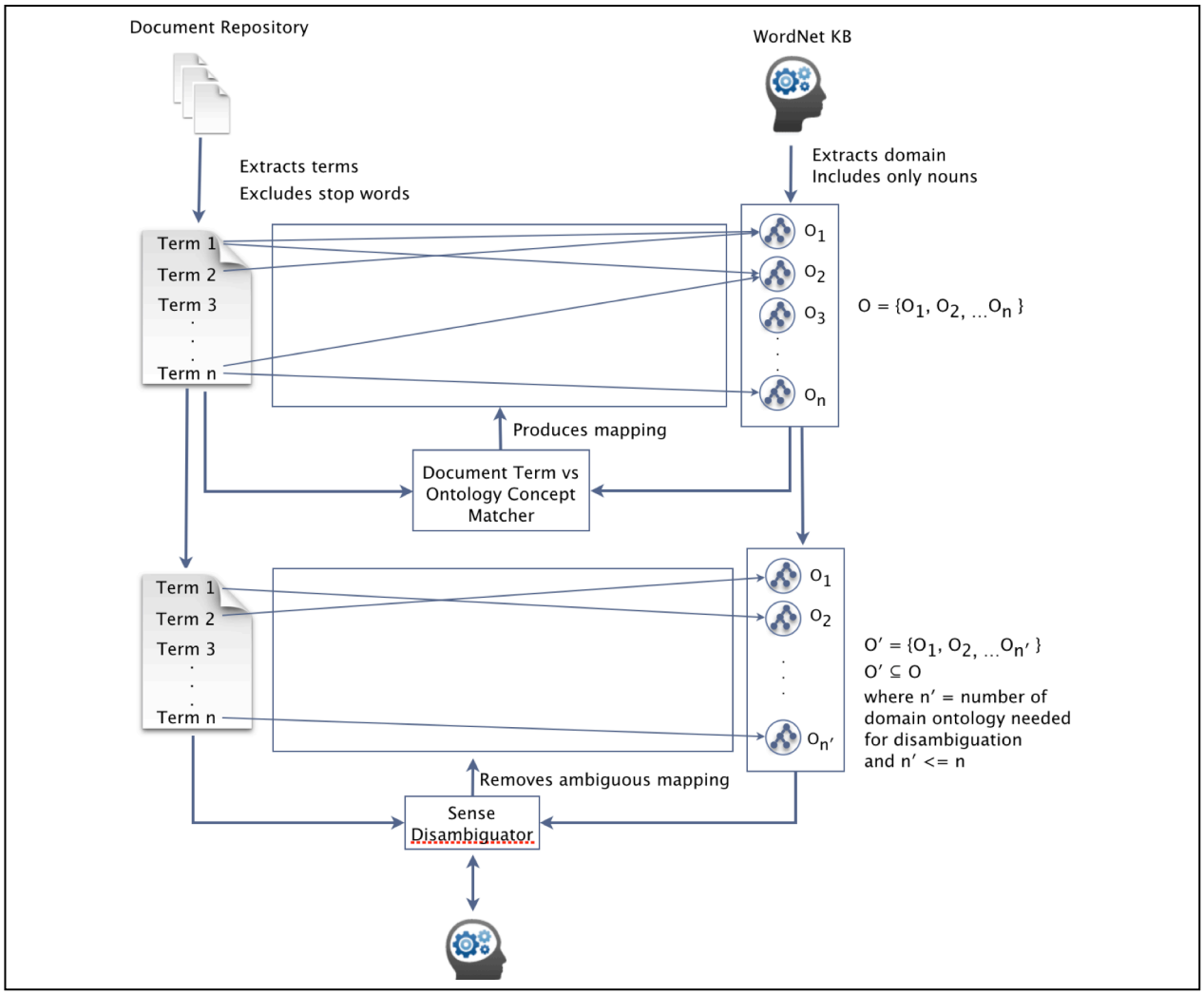

Figure 2 Sense disambiguation methodology

At this stage, it performs sense disambiguation of the terms which fall into multiple domains but do not represent the same concept. For example, the term France appears in the following two domains: organism (a topic domain maintains the writer sense) and France (a region domain maintains the country sense). The sense disambiguation methodology is sketched in Figure 2 and described below.

1. For each term, retrieve all noun synsets and their hierarchies including only hypernyms, hyponyms, holonyms and meronyms from WordNet KB. Starting from the nearest neighbors, check the presence of the terms from the more specific and more generic synsets in the documents' content. In the case of finding a match, prioritise the hierarchies based on the proximity of the matched synset. Select the synset of the hierarchy with the highest proximity.

   In other words, given that there are two synset hierarchies H1 and H2 connected to a term x, where

   H1 = {A1} → {B1} → {C1},
   H2 = {A2} → {B2} → {C2},
   the relational symbol → is read as 'has hypernym',
   the term x appears both in synsets {A1} and {A2},
   a term y is in {B1} and another term z is in {C2}, and
   both y and z are present in the documents' content.

   In this case select the sense attached to the synset {A1} of the hierarchy H1 as the more relevant sense for the term x because the distance between {A1} and {B1}, is 1 hop only, which is lower than the distance between {A2} and {C2}, 2 hops.
   If more hierarchies correspond to the highest proximity or no neighbor is matched, go to Step 2.
2. For a term retrieve the senses from the domain ontologies and sense ranks from WordNet KB. Perform comparison among the sense ranks. The sense with the higher rank is selected as the more relevant sense for the given term. It can happen that the same domain ontology contains two different senses for the same term. In this case the disambiguation is performed using sense rank, similarly as above. If the term appears in two domains with the same sense, leave it with the more relevant domain. If the term is not available in the domain ontologies, go to step 3.
3. Take the highest ranked sense for the term which does not appear in any of the domain ontologies but is present in WordNet KB. The term which is neither present in the domain ontologies nor in WordNet KB is not subject to disambiguation.

Disambiguation is followed by the semantic enrichment of the document terms. In this enrichment, it connects each term with the more generic term residing one level above in the hierarchy and with the semantically equivalent terms, whenever available. Note that for the enrichment it uses both the domain ontologies and WordNet KB. Finally, it produces an ontology called Document Term Ontology (DTO) with the

codification of the terms and their semantics in RDF. More generic or *is-a* subsumption relations are represented as *rdfs:subclassOf* relations. It uses the *owl:equivalentClass* relation for representing the semantically equivalent terms. Here the terms refer to concepts or classes. Terms representing the same entity are codified with *owl:sameAs* relation. The relation between an entity term and a concept term is represented with *rdf:type* relation.

Document term semantification and ontology generation were done completely automatically. The terms and their relations in RDF are codified in order to give them a formal representation and make their semantics recognisable by the Semantic Web tools and technologies such as Jena and SPARQL.

5 Semantic Search

Figure 3(a) demonstrates the faceted smart search system, developed for the EnAlgae project, allowing users to query documents using keywords typed in the search box and/or selecting facets, i.e., document type, year of publication, region, keyword and project action, from the left panel to further narrow down the search query. For each document, project partners have provided us with a list of metadata to fill out the facets. The search was limited to the exact match of all keywords (except the stop words) provided at the search box together or individually with one or more of the metadata fields. Therefore, while a search for *algae cultivation in Netherlands* produces 7 documents, replacing *Netherlands* with *Holland* returns no results, as shown in Figure 3(b). The reason for experiencing an empty result is the absence of *Holland* in the list of metadata provided by the partners for each document. The search can be explored at the following link: https://ixion.bcu.ac.uk/enalgae/facetedSearch. The challenge remained about providing a system that can answer user queries seamlessly when alternative terms are used to refer to the same real world entity.

Semantic search was thought of as a means to overcome this issue (Fernández et al., 2008; Cohen et al., 2003; Kandogan et al., 2006; Hilderbrand et al., 2007; Giunchiglia et al., 2009). It deals with the generation of query responses not only by syntactically matching query terms with the content of the documents but also by taking into account the semantics of both the content and the terms (Uren et al., 2007; Tumer et al., 2009; Madhu et al., 2011; Ferr´e and Hermann, 2011; Chu-Carroll et al., 2006).

It has performed the semantic computation of the documents' terms in advance offline and the results are kept in the DTO ontology. The DTO

ontology contains one meaning per term. Therefore, this ontology is used for the sense disambiguation of the query terms. The rationale behind choosing the sense available in this ontology as the more relevant one is that the query is targeted towards the documents from which the ontology is built. Hence the query terms and the ontology terms are highly likely to share the same meaning.

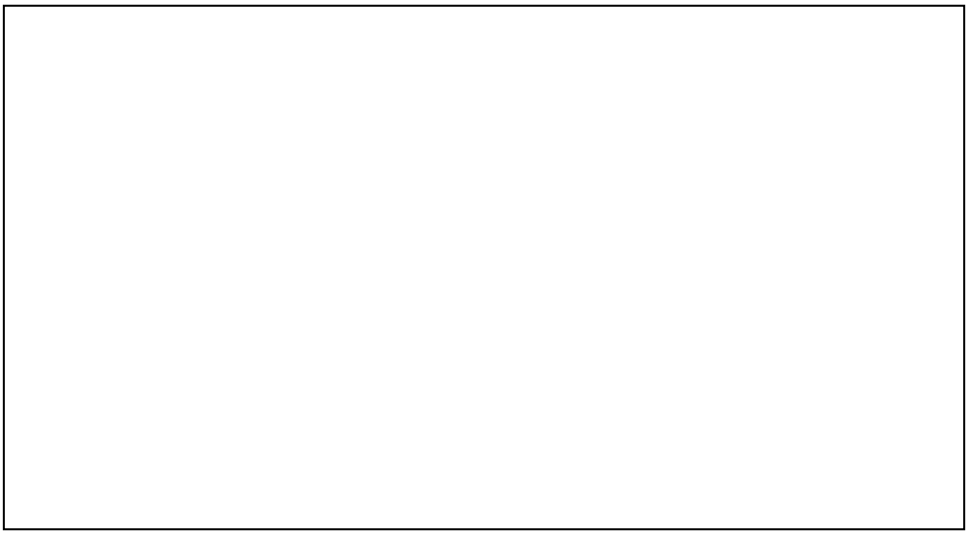

Figure 3(a) Faceted smart search returning 7 documents for algae *cultivation in Netherlands*

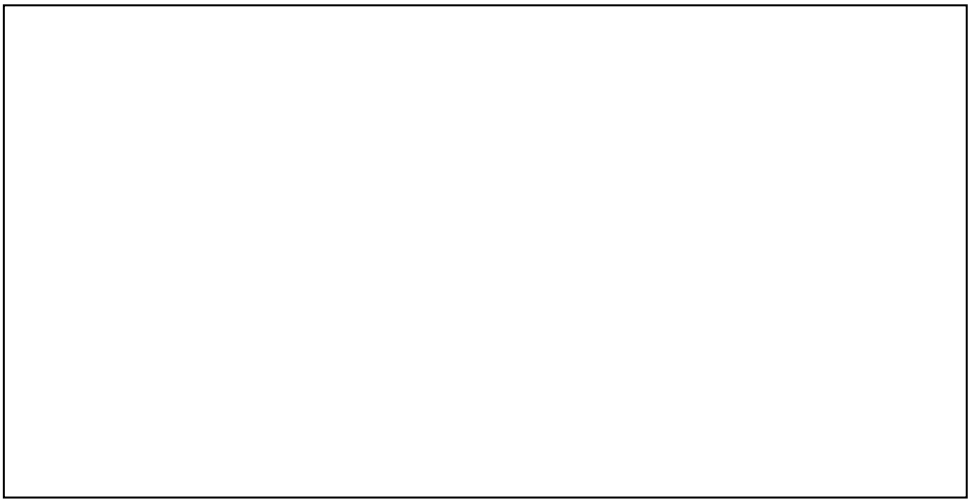

Figure 3(b) Faceted smart search returning no documents for the search *algae cultivation in Holland*

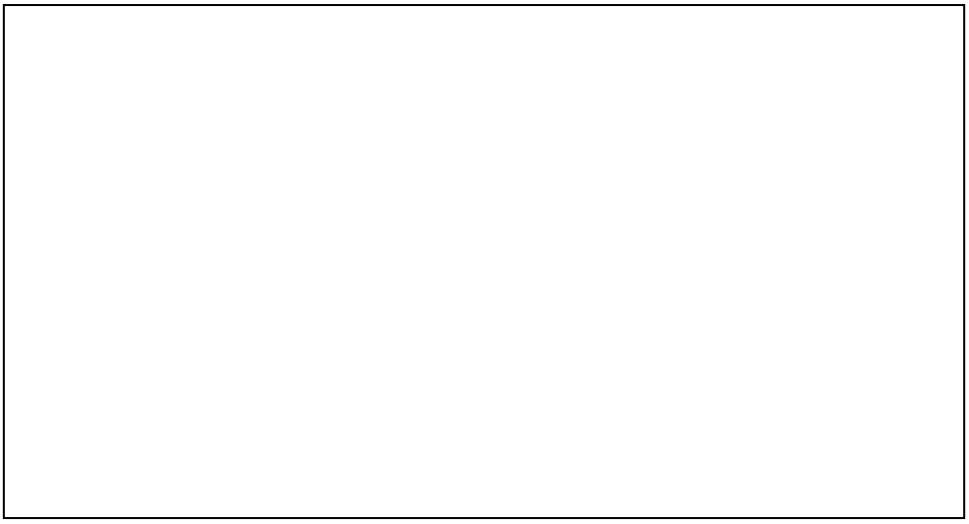

Figure 3(c) Faceted Semantic Search returning 7 documents for the search *algae cultivation in Holland*

Figure 3(c) shows the developed semantic search application that uses the DTO ontology to retrieve any semantically equivalent terms of a query term. It looks for the appearance of them in an inverted index, which keeps track of the mapping between a term and the documents in which it appears. Note that the index might not maintain the mapping for all the semantically equivalent terms. Finding any terms in the index which are semantically equivalent to the query term returns the mapped documents. It can understand that UK and United Kingdom are the same entity and similarly that Holland and Netherlands refer to the same real world entity. It returns the same set of documents for the query *algae cultivation in Holland* and *algae cultivation for Netherlands*. To achieve an acceptable query response time (i.e., less than a second) we have represented the DTO ontology and the index in JSON. It was observed that the response time is satisfactory. The semantic search application can be explored further at the following link: https://ixion.bcu.ac.uk/enalgae/facetedSemanticSearch.

6 Experimental Results and Evaluation

We have conducted experiments with the documents received from the EnAlgae project partners located in the North West European (NWE) region. These documents are mainly project deliverables produced between 2011 and 2015 inclusive. They fall under various categories such as report, policy and factsheet. Reports and factsheets describe, among others, algae growth, cultivation and initiative in the countries of NWE

region. 87 documents in total were received containing around 250,500 terms. The number of unique terms excluding stop words is 14,157.

As shown in Table 1, in WordNet version 2.1, 81,246 noun synsets accumulate 117,097 terms. There are 526 domains and 51,549 synsets. The numbers of domains in topic, region and usage categories are 342, 161 and 23, respectively. Some synsets overlapped in the domains under the top categories. There are 77,575 monosemous (single meaning) terms and 7,510 polysemous (multiple meanings) terms. Average polysemy of all terms is 1.12 and for polysemous terms is 2.39.

Table 1 Statistics about WordNet 2.1

Number of noun synsets	81,246
Number of terms in the noun synsets	117,097
Number of domains	526
Number of synsets in the domains	51,549
Number of domain categories	3
Number of domains in the topic category	342
Number of domains in the region category	161
Number of domains in the usage category	23
Number of monosemous terms in the noun synsets	77,575
Number of polysemous terms in the noun synsets	7,510

As shown in Table 2, out of the 14,157 unique terms in the 87 project documents, 4,160 terms in total were found in WordNet. Out of these, 1,671 terms are monosemous and 2,489 terms are polysemous. In nearly 90% of cases monosemous terms are available with the intended meaning. Out of 526 domains, 342 have been found as relevant. The number of terms available in these domains is 2,689, where 1,595 terms are monosemous and 1,094 terms are polysemous. In the case of polysemous terms within the domains, our disambiguation algorithm showed interesting results.

Table 2 Statistics about the EnAlgae project documents

Number of documents	87
Number of unique terms in the documents	14,157
Number of unique terms available in WordNet	4,160
Number of unique monosemous terms available in WordNet	1,671
Number of unique polysemous terms available in WordNet	2,489
Number of domains relevant to the documents	342

Number of unique terms found in the domains	2,689
Number of unique monosemous terms found in the domains	1,595
Number of unique polysemous terms found in the domains	1,094

By applying semantic enrichment with this disambiguation to the terms of two facets, region and keywords, we have achieved an acceptable accuracy, which helped us develop a usable semantic search application with better performance and user satisfaction.

As reported in Table 3, the query *Eire* (synonym of Ireland) returns 15 documents. All of these documents are relevant to the query. Total number of relevant documents is 16. Therefore, precision = 15/15, recall = 15/16 and f-measure = 2*(15/15)*(15/16)/((15/15)+(15/16)).

Table 3 Precision, recall and f-measure of the queries

Query	Retrieved documents	Retrieved relevant documents	Relevant documents	Precision	Recall	F-measure
Eire	15	15	16	1.0	0.94	0.97
Ireland	20	16	16	0.8	1.0	0.89
Algae cultivation in Eire	6	6	7	1.0	0.86	0.92
United Kingdom	18	17	17	0.94	1.0	0.97
U.K.	27	17	17	0.63	1.0	0.77

The average precision, recall and f-measure of the queries performed and reported in Table 3 are 0.87, 0.96 and 0.90, respectively. The developed domain specific Faceted Semantic Search outperforms the state of the art systems of its kind.

7 Related Work

Clever Search (Kruse et al., 2005) uses WordNet for extending query with semantically equivalent terms. Query term disambiguation is done with user intervention. It allows a user to search with a single word or multiword term only. Similar to this approach, we also use WordNet for retrieving semantics. However, our approach differs in the following ways. While Clever Search allows user intervention for sense disambiguation, our approach accomplishes this completely automatically and allows queries with multiple terms.

Moldovan and Mihalcea (2000) proposed a term sense disambiguation approach that searches the Web with a sequential pair of terms appearing in a query. By replacing one term with a synonym retrieved from WordNet and keeping the other term unchanged and the number of hits are counted. The search is performed iteratively for all possible synonyms. Similar search iterations are also done by altering the previously unaltered term with its synonyms one by one and making the other term static. The synset whose terms contributed to the maximum number of hits is taken as the right sense. It can be argued that this approach is computationally expensive and might fail to respond to user queries in reasonable time. Our approach employs mainly domain knowledge extracted from WordNet for disambiguating query terms.

Conceptual graph matching (Zhong et al., 2002) exploits semantic relations of WordNet to enhance the search by matching terms that are more specific. It converts the set of documents to which the search is applied into concept graphs (CGs). Links are maintained between the CGs and the documents. User query is also converted into a concept graph which is then matched with the CGs in order to return the corresponding documents. However, while generating the CGs from queries, it needs user intervention to specify the entry node of the graph (similar to the root node in a tree). Unlike these approaches, all our computations are performed automatically, leaveing users with no additional burden apart from typing the query.

Buscaldi et al. (2005) demonstrated the use of WordNet to improve search results. They dealt with geographical locations and found that the holonym (part of) relation was more significant for query expansion. The idea behind this was to return documents describing a part (e.g., research in Birmingham) when the query is about its parent (e.g., research in the United Kingdom). In this approach users are asked to disambiguate the query terms, e.g., place name is required to written with prefix geo:. In contrast, our work accomplishes disambiguation automatically.

8 Conclusion and Future Work

A detailed description has been provided about how a traditional keyword search system can be extended with semantic capability. We have described how the term sense disambiguation issue has been partially addressed. Our paper proposes an approach for term sense disambiguation. Finally, we have performed evaluation of the developed semantic search application, which showed favorable outcomes. Our future work will investigate the performance of our sense disambiguation approach by

applying it in other domains such as automotive and aerospace engineering.

References

Auer, S., Dietzold, S. and Riechert. T. (2006) 'OntoWiki – A tool for social, semantic collaboration', *ISWC '05 Proceedings of the 5th Int. Semantic Web Conference*, No. 4273 in LNCS, Springer, pp.736–749.

Bentivogli, L., Forner, P., Magnini, B. and Pianta, E. (2004) 'Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing', *In COLING Workshop on Multilingual Linguistic Resources*, Geneva, Switzerland, pp.101–108.

Bonino, D., Corno, F., Farinetti, L. and Bosca, A. (2004) 'Ontology driven semantic search', *WSEAS Transaction on Information Science and Application*, 1, pp.1597–1605.

Buscaldi, D., Rosso, P. and Sanchis. E. A. (2005) 'A WordNet-based queryexpansion method for geographical information retrieval', *In CLEF Workshop at GeoCLEF*, Vienna, Austria.

Chang, S. F., Ma, W. Y. and Smeulders, A. (2007) 'Recent advances and challenges of semantic image/video search', *Proceedings of IEEE Int. Conf. on Acoust., Speech, Signal Processing*, vol. 4, pp.1205–1208.

Chirita, P. A., Gavriloaie, R., Ghita, S., Nejdl, W. and Paiu, R. (2005) 'Activity based metadata for semantic desktop search', *ESWC '05 Proceedings of the 2nd European Semantic Web Conference*, Heraklion, Greece, pp.439–454.

Chu-Carroll, J., Prager, J., Czuba, K., Ferrucci, D. and Duboue, P. (2006) 'Semantic search via XML fragments: a high-precision approach to IR', *In SIGIR*, pp.445–452.

Cohen, S., Mamou, J. Kanza, Y. and Sagiv, Y. (2003) 'XSEarch: a semantic search engine for XML', *VLDB '03 Proceedings of the 29th international conference on Very large data bases - Volume 29*, pp.45–56.

Dean, M. and Schreiber, G. (Edition) (2004) 'OWL Web Ontology Language Reference', *W3C Recommendation*.

Fang, W., Zhang, L., Wang, Y. and Dong, S. (2005) 'Toward a Semantic Search Engine Based on Ontologies'. *Intl. Conference on Machine Learning and Cybernetics*.

Fernández, M., Sabou, V. M., Uren, V., Vallet, D., Motta, E. and Castells. P. (2008) 'Semantic Search meets the Web', *ICSC '08 Proceedings of the 2nd IEEE International Conference on Semantic Computing*, Santa Clara, CA, USA, pp.253–260.

Ferr´e, S. and Hermann, A. (2011) 'Semantic search: reconciling expressive querying and exploratory search', *ISWC '11 International Semantic Web Conference*, LNCS 7031, Springer, pp.177–192.

Ganbold, A., Farazi, F. and Giunchiglia, F. (2014). 'An Experiment in Managing Language Diversity Across Cultures', *Proceedings of the Sixth International Conference on Information, Process, and Knowledge Management*.

Giunchiglia, F., Kharkevich, U. and Zaihrayeu, I. (2009) 'Concept Search', *ESWC '09, Proceedings of the 6th European Semantic Web Conference*, Heraklion, Greece, pp. 429–444.

Giunchiglia, F., Maltese, V. Farazi, F. and Dutta, B. (2010) 'GeoWordNet: A Resource for Geo-spatial Applications', *ESWC '10 Proceedings of the 7th Extended Semantic Web Conference*, Heraklion, Greece, pp.121–136.

Guha, R., McCool, R. and Miller, E. (2003) 'Semantic search', *WWW '03 Proceedings of the 12th international conference on World Wide Web*, ACM, New York, NY, USA, pp.700–709.

Haase, P., Herzig, D. M., Musen, M. and Tran. D. T. (2008) 'Semantic Wiki Search', *Tech. rep.*. URL: http://www.aifb.uni-karlsruhe.de/WBS/pha/publications/semwikisearch.pdf.

Hildebrand, M., Ossenbruggen, J. and van Hardman, L. (2007) 'An Analysis of Search-based User Interaction on the Semantic Web', *Information Systems Journal*, INS - E0706, CWI, Amsterdam, Holland, pp.1386–3681.

Hyvonen, E., Saarela, S. and Viljanen. K. (2004) 'Application of ontology-based techniques to view-based semantic search and browsing', *ESWS '04 The semantic web: research and applications*, Heraklion, Greece. Springer–Verlag, Berlin, pp.92–106.

Kandogan, E., Krishnamurthy, R., Raghavan, S. Vaithyanathan, S. and Zhu, H. (2006) 'Avatar semantic search: a database approach to information retrieval', *Proceedings of the ACM SIGMOD international conference on Management of data*, New York, NY, USA, pp.790–792.

Kruse, P. M., Naujoks, A., Roesner, D. and Kunze, M. (2005) 'Clever search: A wordnet based wrapper for internet search engines', *Proceedings of the 2nd GermaNet Workshop*.

Lei, Y., Uren, V. and Motta, E. (2006) 'Semsearch: A search engine for the semantic web' *EKAW '06 Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management*, pp.238–245.

Madhu, G., Govardhan, A. and Rajinikanth, T. (2011) 'Intelligent Semantic Web Search Engines: A Brief Survey', *International journal of Web & Semantic Technology*, 2(1), 2011, pp.34–42.

Maltese, V. and Farazi, F. (2011) 'Towards the Integration of Knowledge Organization Systems with the Linked Data Cloud', *UDC seminar*.

Mangold, C. (2007) 'A survey and classification of semantic search approaches'. *Int. J. Metadata Semantics and Ontology*, 2(1), pp.23–34.

Miller, G. (1995) 'WordNet: A lexical database for english', *CACM '95*, 38(11):39–41.

Moldovan, D. I. and Mihalcea, R. (2000) 'Using WordNet and lexical operators to improve internet searches', *IEEE Internet Computing*, 4(1), pp.34–43.

Sapkota, K., Raju, P., Byrne, W. and Chapman, C. (2015) 'Semantic Economic Models for Bioenergy Projects', *International Journal of Semantic Computing*, Vol. 9, No. 3, pp.333–352.

Schaffert, S. (2006) 'IkeWiki: A Semantic Wiki for Collaborative Knowledge Management', *STICA '06 Proceedings of the 1st International Workshop on Semantic Technologies in Collaborative Applications*.

Schreiber, G., Amin, A., Aroyo, L., van Assem, M. V., de Boer, V., Hardman, L., Hildebrand, M., Omelayenko, B. van Osenbruggen, J. Tordai, A., Wielemaker, J. and Wielinga, B. (2008) 'Semantic annotation and search of cultural - heritage collections: The Multimedian e-culture demonstrator', *J. Web Semantics*, Vol. 6(4), pp.243–249.

Srinivasan, N., Paolucci, M. and Sycara, K. (2004) 'An efficient algorithm for OWL-S based semantic search in UDDI', *Proceedings of SWSWPC at the 2nd ICWS*, San Diego, CA, USA, pp.96–110.

Suchanek, F. M., Kasneci, G. and Weikum, G. (2007) 'Yago: A Core of Semantic Knowledge', *Proceedings of the 16th international World Wide Web conference*, New York, NY, USA.

Tumer, D., Shah, M. A. and Bitirim, Y. (2009) 'An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia', *In: Fourth International Conference on Internet Monitoring and Protection*.

Uren, V., Lei, Y., Lopez, V., Liu, H., Motta, E. and Giordanino, M. (2007) 'The usability of semantic search tools', *A review. Knowl. Eng. Rev.*, 22(4), pp.361–377.

Zhong, J., Zhu, H., Li, J. and Yu, Y. (2002) 'Conceptual Graph Matching for Semantic Search', ICCS, pp.92–106.

Kaniovskyi, Y., Benkner, S., Borckholder, C., Wood, S., Nowakowski, P., Saglimbeni, A., and Lobo, T. P. (2015) 'A Semantic Cloud Infrastructure for Data-intensive Medical Research. International Journal of Big Data Intelligence, 2(2), pp.91–105.