Using Spreading Activation to Evaluate and Improve Ontologies

Rónan Mac an tSaoir
Watson Group, IBM Ireland
ronanate@ie.ibm.com

Abstract

In this paper, we explore the relationship between the human-encoded semantics of ontologies and their application to natural language processing (NLP) tasks, such as word-sense disambiguation (WSD), for which such ontologies may not have been originally designed. We present a method for assessing the semantic content of an ontology with respect to a target domain, by spreading activation over a graph that represents instances of ontology concepts and relationships, in domain text. Our proposed method has several advantages beyond existing ontology metrics. By identifying bias or imbalance in the ontology, we can suggest target areas for improvement, and simultaneously facilitate the automated optimisation of the graph for use in the chosen NLP task. On applying this method to the Unified Medical Language System (UMLS) ontology, we significantly outperformed existing graph-based methods for WSD in biomedical NLP (0.82 accuracy). The subsequent introduction of a fall-back mechanism, using word-sense probability, achieved state of the art for unsupervised biomedical WSD (0.89 accuracy).

1 Introduction

Although ontologies do encode human knowledge, the degree to which these artefacts represent the entire scope of semantics in a target domain is difficult to quantify. Since few ontologies offer large enough scope to cater for an entire domain in natural language, merging of multiple ontologies is often necessary (Noy, 2004). This further compounds the problem of assessing the semantic relevance of the merged resource. The collective semantics in multiple source ontologies can often overlap inconsistently, and negotiation of meaning so that the associated set of concepts and relationships in the ontology remains balanced, is critical. The merging process is usually reserved for domain experts, who focus on ontology portions in which they specialise. It's generally a case of painstakingly mapping individual concepts between component data sets, to ensure semantic integrity (Jiménez et al, 2012). Coordinating collaborative ontology editing and merging is a related and well-known problem (Jiménez et al, 2011).

Existing ontology metrics generally focus on structural and logical semantics (Sicilia et al, 2012). Assessing how closely ontologies match the semantics of natural language text, or identifying specific portions of an ontology which require further development, are more difficult tasks. We have identified a robust method for this assessment. This method involves static analysis of a graph representing ontology instances and inter-concept relationships, to address apparent imbalances that hinder spreading activation in the graph. When accuracy and relevance for the task improves, the modified graph or activation strategy identifies portions of interest for further development. Many ontologies used in NLP today are not designed for this (Guarino et al, 2009), and a flexible, automatic evaluation method is useful.

We focused on the Unified Medical Language System (UMLS) as a typical ontology (NLM, 2013), displaying many of the problems associated with use of ontologies in NLP, including merged terminology, strongly overlapping semantic categories, inconsistent levels of structural depth, as well as inconsistent coverage of associated instance data (Pisanelli et al, 1998). We chose to assess this ontology with respect to word sense disambiguation (WSD), which is commonly accepted to be one of the most difficult tasks in NLP (Navigli, 2009). We used the MSH-WSD corpus for testing purposes, which commonly used in assessing methods for biomedical WSD (Jimeno Yepes and Aronson, 2012; McInnes et al, 2011; Gad el Rab et al, 2013). Using node-centric graph metrics, we identified portions of the ontology which were not conducive to WSD via spreading activation. After appropriately modifying the activation strategy, we achieved state of the art performance in graph-based biomedical WSD (0.82).

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2237–2248, Dublin, Ireland, August 23-29 2014.

2 Background

2.1 Ontologies

An ontology, in computer science, is defined as an 'explicit specification of a shared conceptualization' (Gruber, 1993), where a conceptualization may be some subset of real-world semantics, with respect to the requirements for a given task. It can contain concepts or classes of object, object properties, and inter-concept relationships, as well as instances of these in the target domain. Such structured resources facilitate the sharing and re-use of domain knowledge, and are invaluable for NLP applications. A primary example of such a resource is the UMLS, provided by the National Library of Medicine (NLM, 2013). The data set consists of a large lexicon, including millions of instance surface forms, in conjunction with an ontology of concepts and inter-concept relationships in the medical domain. It is composed of 139 different source ontologies or terminologies, each of which have their own labels, descriptions and semantic perspective (e.g. FMA[2] for the body, and RXNORM[3] for drugs, as well as more general ontologies like SNOMED[4]). An example ontology is shown in Figure 1 below.

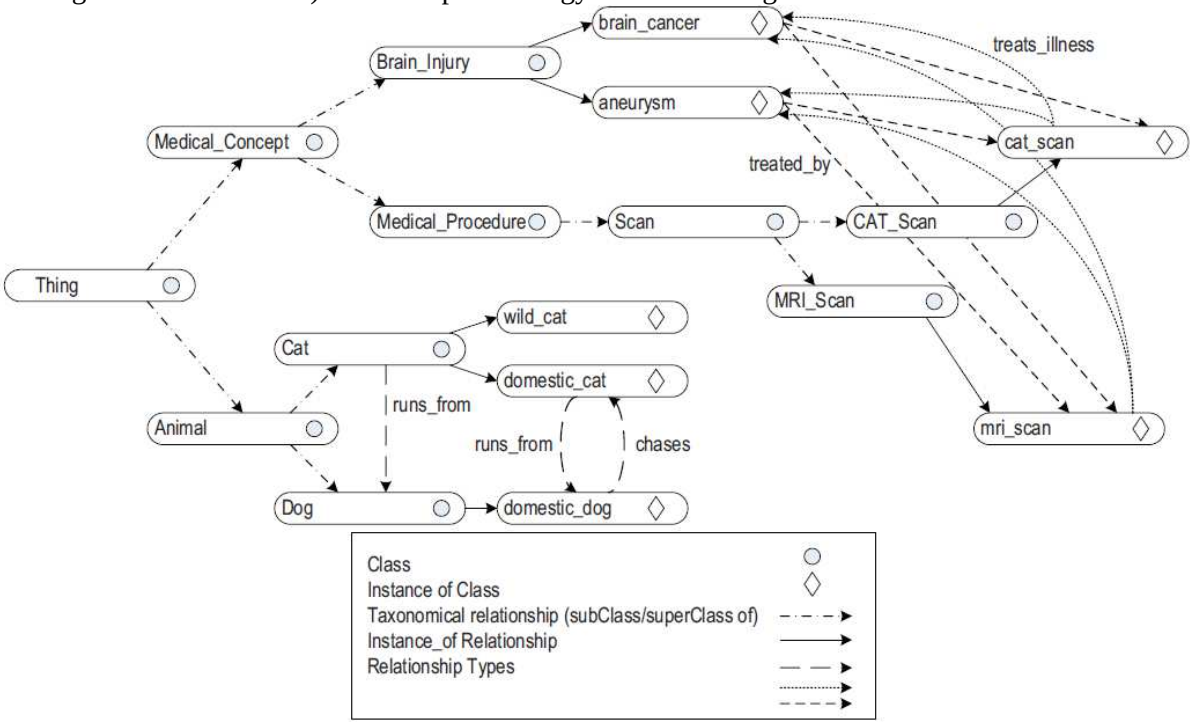

Figure 1. A simple (and incomplete) ontology describing ambiguous senses of the word "cat"

2.2 Ontology Evaluation

Evaluating ontology semantics commonly focuses on the structural and logical nature of the resource. Related efforts may use logical reasoning to ensure that the semantics are internally consistent, or use the structure and labels of another ontology as a baseline, assuming that textual labels for synonymous concepts will be consistent between sources (Vrandecic and Sure, 2007; Ma, 2013). A metric which goes beyond these and evaluates the semantic relevance to a given task is sorely needed (Vrandecic and Sure, 2007). While metrics that examine the completeness of an ontology's content are suggested in the literature (Tartir et al, 2005), these metrics reflect a high-level summary of the content. The evaluation of this content, independent of the ontology itself, and at a sufficiently fine-grained level to suggest areas for improvement, would be of significant additional benefit.

Vrandecic and Sure (2007) recognise the paucity of metrics that take the ontology semantics into account. In terms of semantic quality, they propose leveraging a logical reasoner to evaluate that an ontology is consistent within the context of its own assertions. However, there is no objective analysis of the semantic content with respect to real world human knowledge. Ma et al (2013) point out that prior

[2] FMA: http://sig.biostr.washington.edu/projects/fm/AboutFM.html
[3] RXNORM: http://bioportal.bioontology.org/ontologies/RXNORM
[4] SNOMED: http://bioportal.bioontology.org/ontologies/SNOMEDCT

ontology metrics neglect implicit semantic knowledge. They acknowledge the utility of a graph structure in representing the content of an ontology, and assert that this structure preserves well the semantics of the ontology. However, they do not proceed to examine the ontology in the context of a real-world semantic evaluation. By limiting the scope of comparison to sets of related ontologies, they work on the assumption that similarly labelled concepts and structures are roughly equivalent. Additionally, Sicilia et al (2012) suggest that there is no obvious metric to identify when an ontology needs to be improved.

We propose that graphs composed of instances of ontology concepts and relationships, along with associated unique identifiers, are a less naïve approach to semantic matching than textual labels. We suggest an objective analysis of how annotated instances of ontology concepts and relationships interact, by a process such as spreading activation in an associated graph, would be more reflective of the proximity of the evaluated ontology to the semantics of the target domain text. We also suggest that analysis of particular characteristics of the graph, that amplify or hinder this activation process, are helpful in identifying specific portions of the associated ontology that require further development. Interestingly, the use of spreading activation as a method for ontology assessment has already been carried out previously (Fang and Evermann, 2010). In that case however, the spreading activation was in the context of cognitive psychology, where test subjects manually assessed ontology content. An automated approach, leveraging the same principles, without the requirement for human reviewers, would be of great value.

2.3 Word Sense Disambiguation

WSD is one of the most critical tasks in NLP (Navigli, 2009), and is often described as AI complete. Navigli (2009) identifies several main categories of approach to WSD, namely knowledge based, supervised and unsupervised methods. He proposes knowledge based methods as the most useful in the medium to long term, for several reasons. He points to the availability of knowledge resources such as WordNet, Yago, and DBPedia, resources which are actively developed and enriched, as a starting point of significant value. He also suggests that supervised approaches are better for categorisation tasks like part-of-speech (POS) tagging, rather than tasks that require more fine grained detail such as real-world word-sense disambiguation. As an example of this, consider that the process of disambiguating the correct POS for a word may involve the selection of one from a set of possible POS tags. One such tagset, widely used for English, is the Penn Treebank tagset consisting of 36 separate tags. The UMLS data set, however, contains close to 3 million[5] distinct senses.

Though WSD is still widely regarded as an unsolved problem, supervised approaches to WSD generally perform well. Navigli (2009) suggests that this is due to the lack of real-world considerations in development and testing of WSD methods. We can consider the MSH-WSD corpus as an example demonstrating typical limitations when compared with the requirements for a real-world system. MSH-WSD is a commonly used data set in biomedical WSD, using sense IDs from UMLS, and consisting of approximately 37,000 separate documents or abstracts, where a single ambiguous sense is annotated with the correct UMLS sense ID. A WSD system need only identify this single sense correctly (regardless of the other words in the document), in order to score highly. Additionally, there are a total of 423 distinct word-senses annotated in this test set, greatly reducing the scope of the task involved from approximately 3 million possible senses in the full UMLS. As a result, this data set is not a strong reflection of what is required in real-world biomedical NLP applications, where a high percentage of the words in a given document or context must be assigned their correct senses.

It is generally accepted that unsupervised methods for WSD minimise the cost of developing a suitable application, by relying on features that may be extracted directly from the target domain text, or alternatively using existing knowledge in some form. The latter are often referred to as knowledge-based (KB) methods. For supervised WSD a gold-standard is required input, where manually curated data sets facilitate the training of robust machine learning algorithms. Supervised methods generally outperform unsupervised (Agirre et al, 2010), but are limited by the cost of developing the required training data. However, as mentioned previously, these systems may not perform so well in real world WSD scenarios.

In a biomedical context, there are several examples of both supervised and unsupervised (including knowledge-based) approaches. Most unsupervised approaches leverage the UMLS to some extent, and build on that knowledge using methods like Automated Corpus Extraction (Jimeno Yepes and Aronson, 2012) and Information Content Similarity (McInnes et al, 2011). The commonly cited example of a

[5] UMLS stats: http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html

supervised approach that consistently outperforms known unsupervised approaches is Naïve Bayes (Jimeno Yepes and Aronson, 2012; McInnes et al, 2011), achieving 0.94 accuracy on the common data set, although as we've outlined previously, the search space for a correct tag in the chosen data set (with a total number of 423 senses) is much smaller than would be the case in a real-world system.

Several recent approaches to biomedical WSD leverage structured knowledge in the form of a graph. Examples range from the use of co-occurrence data from a domain-specific corpus (Agirre et al, 2006), to variations of PageRank (Agirre and Soroa, 2009; Agirre et al, 2010), to the representation of an ontology, or portion of an ontology, as a graph (Gad El-Rab et al, 2013). Ontologies are often used as a source from which to build the required graph, as they are readily available in many domains, and provide a starting point of high-quality semantic knowledge. As identified previously, the lexical ontology Wordnet is a commonly used resource in open domain WSD. Similarly, in the biomedical domain, the UMLS is equally common. Hybrid approaches leveraging both general lexical semantics like WordNet with domain-specific semantics like UMLS are not as common however, but have been used with promising results in other related NLP tasks such as anaphora resolution (Liang and Lin, 2005).

Graph based methods have not performed as well as other unsupervised approaches, like Machine Readable Dictionaries: 0.8070 (Jimeno Yepes and Aronson, 2012), semi-supervised Automated Corpus Extraction methods: 0.8383 (Jimeno Yepes and Aronson, 2012), and co-occurrence metrics: 0.78 (McInnes and Pedersen, 2013). A recent approach (El-Rab et al, 2013) achieved mixed results with respect to particular terms in the MSH-WSD test corpus, achieving an overall accuracy of 0.603. State of the art accuracy for graph-based methods, in unsupervised biomedical WSD, was 0.72 (McInnes et al, 2011). State of the art in overall unsupervised biomedical WSD was 0.87 (Jimeno-Yepes and Aronson, 2012).

2.4 Spreading Activation

The theory of spreading activation was first proposed by Quillian (1966), in a model of human semantic memory. Quillian proposed an abstract model of human memory, in order to artificially represent the means by which a human's brain might process and understand the semantics of natural language. This model was enhanced by Collins and Quillian (1969) for retrieval tasks, and further modified by Collins and Loftus (1975). The latter provided inspiration for research in many other related fields, from cognitive psychology to neuroscience, to natural language processing, among others (Pace-Sigge, 2013).

The basic premise of spreading activation is related to that of connectionism in artificial intelligence, which uses similar models for neural networks to reflect the fan-out effect of electrical signal in the human brain. In the case of neural networks, a vertex in the graph could represent a single neuron, and edges could represent synapses. In information retrieval (Crestani, 1997) and word-sense disambiguation (Tsatsaronis et al, 2007), generally vertices will represent word-senses and edges will represent some form of relationship, either lexical or semantic linkage, between these senses.

An example implementation is 'Galaxy', developed as part of the Nepomuk Social Semantic Desktop[6], which uses spreading activation to perform clustering on a graph. Instead of traditional methods of hard clustering, which partition a graph into different groups, Galaxy performs soft clustering, which involves identifying a sub-graph located around a set of input nodes, and then finding the focus of this sub-graph. The same implementation provides a configurable weighting model that allows modification of starting weights associated with semantic types, edges and individual nodes in the graph. This has already been used in various scenarios, such as social network analysis and dynamic semantic publication of web content[7], and may also be applied to any set of graph-structured data (Troussov et al, 2008).

By discovering instances of ontology concepts in domain text, using the set of unique identifiers for instances, we can activate corresponding nodes in the graph, from where a signal will traverse outward across adjacent nodes, activating these in turn. As the signal spreads farther from a source node, it gets weaker by an amount specified in an associated weighting model for nodes and edges in the graph. If the signal spreads from multiple nearby source nodes, the signal will combine, and points of overlap will be activated to a greater degree. The nodes which accumulate the most activation are deemed to be the focus nodes for the context. The resulting activated portion of the graph will reflect the inherent meaning of the document, in so far as the ontology's defined semantics will allow.

[6] http://dev.nepomuk.semanticdesktop.org/wiki/TextAnalytics#IBM
[7] http://www.bbc.co.uk/blogs/bbcinternet/2010/07/bbc_world_cup_2010_dynamic_sem.html

To demonstrate this process in action, we will draw examples from the ontology previously defined above. Figure 2 describes the resulting instance graph for the ontology described in Figure 1, on which we can perform spreading activation using instances in text. Firstly, consider the set of surface forms associated with concept instances in table 1. If we annotate the set of contexts below with this lexicon, we can then use the annotations to activate the graph. Nodes that are well connected may benefit from the potential overlap of signal coming from other adjacent nodes. Instances are *italicised* below.

- The *cats* result for the patient's *brain tumour* was assessed by the Doctor.
- *Tigers* and *lions* are *cats* that live in the wild. These *cats* are not afraid of *dogs*.
- The patient survived the *brain tumour*, but died of an allergic reaction to their neighbour's *cats*.

In each example, the ambiguous term is the word “cats”, which can variously refer to: cat_scan, wild_cat and domestic_cat. The surrounding context of each instance contains other concept instances that may help to disambiguate the correct sense of “cats”. In the first example, the nodes representing wild_cat, domestic_cat, cat_scan and brain_cancer will be activated. Since brain_cancer and cat_scan are relatively well connected in the graph, and are also adjacent to one another, the spreading activation will return these nodes as the most likely interpretation of the content.

In the second example, the correct instance is wild_cat. However, this node is isolated in the graph, since there were no associated relationships in the ontology linking this particular instance to other nodes. Since the instance of the class Dog is connected to domestic_cat, these nodes may amplify each other’s signal to a greater degree than is possible at the isolated node wild_cat. It is therefore likely that unless the weighting model is reconfigured, we are unlikely to obtain the correct output. The relevance of isolated nodes may be boosted by increasing the rate of signal decay on other nodes in the graph. However, there is a risk in doing so, since the connectedness of instances in the ontology is likely a better reflection of the semantic content. It would be better to suggest that the ontology would benefit from further development, for example to introduce the ideas of habitat or fear.

The final example demonstrates a more subtle bias in the ontology’s semantics, and the corresponding graph. The overlapping signal from cat_scan and brain_cancer suggests that cat_scan will be returned instead of domestic_cat. Resolving this ambiguity in the graph may require modification of the weighting model, or further development. An advantage in this case however, is the different semantic categories involved: the classes of Cat and Scan. Re-weighting the starting activation signal on the basis of a semantic category is less risky than re-weighting the entire set of nodes in the graph. Even so, further development of the ontology, e.g. to introduce the idea of animal allergies, would be beneficial.

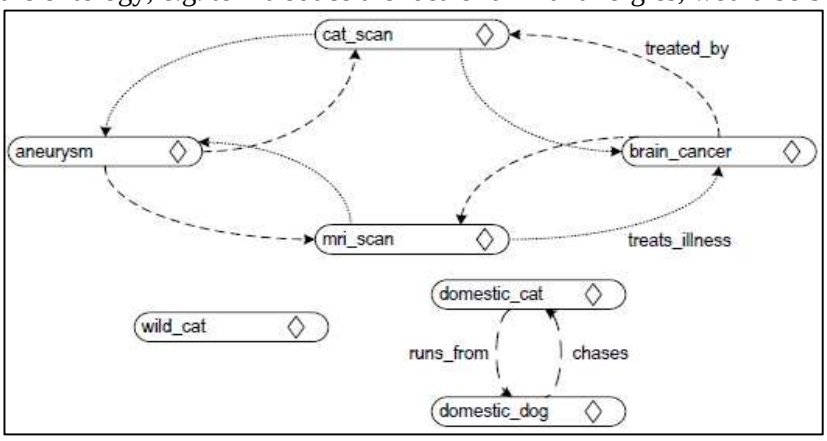

Figure 2. Graph representation of the sample ontology.

Instance ID	Associated Surface Forms
wild_cat	{lions, tigers, cat, cats, cub}
domestic_cat	{cat, cats, kitten }
domestic_dog	{dog, dogs, puppy}
brain_cancer	{brain carcinoma, brain tumour}
cat_scan	{cat, cats, cat scan}

Table 1. Example surface forms for instance data.

3 Method

3.1 Ontology Instance Graph

We extracted data from the UMLS Metathesaurus (MT) and Semantic Network (SN) and built a triple store in RDF/XML[8] format, defining owl:Class and owl:ObjectProperty to reflect concepts and relationships. Using the Galaxy API described in section 2.4, we built a spreading activation network, i.e. a directed graph between instance IDs (vertices) and associated relationships (edges). In order to narrow the proximity between the semantics of domain text and the chosen ontology, we chose to build a graph of instance data. The SN is a high level ontology, and therefore to assume that all relationships between Classes are applicable to all instances would have produced many incorrect assertions, such as "All Drugs have the set of All Drugs as ingredients". Therefore, only instances of relationships that explicitly linked individual concept IDs (Concept Unique Identifiers, CUIs) were used. Across the entire SN, a single CUI may have various types of semantic interactions with other nodes, for example in the context of Drugs and treated Diseases, or separately, in the context of Chemicals and associated Compounds. The UMLS CUIs were used as instance IDs to link surface forms in the text to nodes in the graph.

It is important to point out that the UMLS ontology by no means uses the full expressivity of OWL. However, the general use of spreading activation over a graph derived from ontology content, is not so limited. In other domains, and for ontologies that use the full range of OWL expression, as long as the graph is built from a source that expresses other semantic qualities (e.g. cardinality), the spreading activation strategy will still apply. For example, in the context of our sample ontology, consider activating "cat", the signal spreading to an additional adjacent node for the concept of "four legs", and then other concepts with four legs, such as "dog", becoming activated. The Galaxy API fully supports this.

3.2 Test Corpus and Metric Calculation

We chose to use the MSH-WSD test corpus as our gold-standard. This is a common test set used across the literature in biomedical WSD. The metrics we used were Precision, Recall, FMeasure and Accuracy, whereas prior research mainly focuses on Accuracy. In WSD, a true positive is a disambiguated output that matches a gold-standard, and a false positive is output that does not match. As traditional WSD algorithms are designed to generate output for every word in the text, recall and precision are the same value. However, our algorithm works on the principle of semantic relevance, and there is no guaranteed output; senses with sufficient weight after spreading activation will be displayed. Therefore, we have chosen to take a closer look at precision and recall, which is discussed in more detail in section 4.

Prior literature in biomedical WSD uses older versions of UMLS data, e.g. 2009AB (McInnes et al, 2011). We chose to focus on the 2013AA release of UMLS, in order to assess the most recent version of the ontology's semantic content, and in order to facilitate a useful modification of the current data, which could be leveraged by contemporary NLP systems. This affected the comparison of test results using the MSH-WSD data set.

3.3 Lexical Annotation

In conjunction with the graph described above, we constructed a set of lexical dictionaries that linked UMLS CUIs or instance IDs, to portions of text in a document. These portions of text, otherwise known as surface forms, consisted of potentially many different strings associated with each ID. An example of a data entry for a single UMLS CUI is in table 2 below. Dictionaries were compiled for each semantic category in the UMLS SN, with overlapping associations between ID and textual surface form.

CUI	Semantic Type	Surface Form (Text)
C0018787	BodyPartOrRegion	heart
		cardiac structure
		heart structure
		coronary
		four chambered heart
		the human heart

Table 2. Surface forms associated with the concept "Heart", UMLS CUI: C0018787.

[8] http://www.w3.org/TR/rdf-syntax-grammar/

In order to maximise the potential for spreading activation across the graph, we performed several modifications to the underlying lexical data in UMLS MT, to increase the variations of surface form associated with instances of concepts. Our reasoning for this is as follows: the more instances of concepts that occur in the text, the more nodes that get activated in the graph, and consequently the more opportunities for the activation method to spread out and activate the set of concepts most relevant to the semantics of the document text. For a simple example of this process, please see section 2.4. Examples of transformations carried out in the data are presented in table 3, below.

Pre-existing Term	Transformation Type	New Alternate Surface Form
leg, right	Alternating Comma	right leg
brain cancer	Noun Phrase	cancer of the brain
CANCER	Casing Variants	Cancer
Anaemia	Spelling Variants	Anæmia
Immunoglobulin g	Acronym	Ig
Immunoglobulin g	Term + Acronym	Immunoglobulin g (Ig)

Table 3. Examples of UMLS data transformations applied.

The use of a lexical part-of-speech tagger was particularly effective in filtering out instances of concepts that were obviously introducing unhelpful noise. Some exemplary cases were the Amino Acids "on", "at" and "in" (prepositions), and the GeneOrGenome "was" (verb). UMLS concepts that directly overlapped with words that did not display an appropriate part-of-speech for a true concept (such as adjective or noun), were removed from the document metadata, and thereby not considered as input for spreading activation. For this POS Filter, we chose to use the MaxEntropy model from OpenNLP[9].

3.4 Spreading Activation Strategy

The initial activation strategy was to set starting weights for all semantic categories to a value of 1. Decay factor of the spreading signal at each node in the graph was set to an initial value of 0.5, when the graph was built. The initial threshold of semantic relevance was set to 0.1, and instances retaining a semantic value higher than this would be considered relevant. The lexical annotations from the previous step were used as input to the activation process, and nodes in the graph from instances in the text were assigned their starting weight, according to the number of semantic categories, and their associated weights. As the signal is spread from these starting nodes, the decay factor is applied, reducing the signal strength. For each successive node, the signal is similarly reduced until it falls below the specified threshold, and the activation process is completed. It is important to note that the ambiguity in word-senses may not be entirely removed once the spreading activation has finished. The consequences of this will depend on the particular end-goal. In the case of WSD, we are only interested in obtaining a single most appropriate CUI for a given surface form. We therefore kept only the highest weighted CUI in our system output. In the context of other NLP tasks, such as for named-entity inference or question answering and hypothesis generation (Ferucci et al, 2011), it can be useful to preserve multiple ambiguous outputs for later processing.

It was clear from the outset that simply building a graph of the ontology instance data and semantic relationships was not sufficient to score highly in the WSD task. El-Rab et al (2013), who used the UMLS SN structure for graph-based WSD, reported an overall accuracy of (0.603) on the MSH-WSD test set, which roughly correlates with our baseline system (0.62). Our added advantage is that modification of the weighting strategy allows us to iron out imbalance, or to reduce the influence of those portions of the graph that do not appear to encourage a spreading signal. By focusing on signal amplification and decay, rather than modifying graph semantics, we can change the relevance of particular portions of the ontology without losing any of the original semantic detail. Such modifications are sensitive to performance in the NLP task but, critically, do not require the assistance of domain experts.

We initially pursued a cautious approach to modifying the activation strategy, by only decreasing the starting weight of semantic categories associated with the affected nodes. This weight was decreased by a factor equivalent to the number of overlapping semantic types on the same node. Following this, we measured the accuracy of the approach against the MSH-WSD test corpus for WSD, testing blind, that is by only considering the overall accuracy. Upon close examination of the instance graph, for types of

[9] http://opennlp.apache.org/

structure or characteristics of nodes that may be hindering or over-amplifying the spreading signal (see section 4.1), we further modified the activation strategy to negate the potential influence that certain obviously problematic nodes may have. Modifying our spreading activation strategy in this way, after static graph analysis alone, produced much more accurate output (see table 5, experiment 3).

We then decided to split the test set in the ratio of 4:1, in order to more closely inspect the accuracy of particular cases of WSD, and attempt to correct this specific imbalance in the graph, while still performing some independent validation of the output. The random nature of the split was to choose every fifth example in the data, from the subset for each term. After performing WSD using this 80%, or train set, we discovered that it was possible to distinguish groups of high and low performing nodes in the graph, with respect to the set of static graph metrics, described in the following section.

3.5 Static Graph Analysis (SGA)

As shown in the simple example in 2.4, assessment of ontology semantics can be done up front, before the graph is used. Certain node characteristics may be examined in the graph using a set of graph theoretical metrics, and portions of the graph that are not conducive to spreading activation may be identified. This analysis allows us to make educated modifications to the weighting strategy for spreading activation, as described previously. The set of graph metrics we used is presented in table 4 below.

Metric	**Evaluation**
In Degree	# of inward semantic links
Out Degree	# of outward semantic links
Total Degree	(indegree + outdegree)
Inward Edge Type Variation (ETV)	# of inward edge types
Outward ETV	# of outward edge types
Total ETV	(Inward ETV + Outward ETV)

Table 4. Static Graph Metrics derived from Diestel (2010).

Following the use of these metrics, and the gathering of associated statistics, we categorised particular groups of node in order to apply a common weighting strategy that should maximise performance of the spreading activation algorithm. There were several common patterns that we identified, and chose to target for re-weight. Examples of those nodes that might negatively affect spreading activation are:

- Isolated Nodes, where Total Degree is 0
- Unbalanced Nodes, where inDegree and outDegree are significantly different
- Nodes with few variations in link type, or low Total ETV
- 'Black Hole' nodes, where there is a high Degree to ETV ratio (see section 4.1)

For isolated nodes, we examined the set of associated semantic categories, and boosted their starting weight. For unbalanced nodes, where the indegree was significantly higher or lower than the outdegree, we increased or decreased the decay factor accordingly, to reduce the imbalance of the spreading signal. For nodes with low ETV but high Degree, we increased the decay factor, in order to reduce the potential influence of a single over-used semantic link. For overly promiscuous (Norvig, 1986) or 'Black Hole' nodes, we reduced the starting weight applied by the associated semantic categories, and increased the rate of decay. In certain cases, the intended modifications were incompatible, and resulted in conflicting changes to the graph and weighting strategy. Where certain nodes might require a boost from one category, the starting weight for the same category may need to be reduced, due to an overly-connected node elsewhere. We decided to inhibit the negatively connected nodes only, in light of the increase in system accuracy from reducing noise compared to the gain from improvement of individual nodes.

4 Results and Discussion

The baseline activation strategy was promising. The introduction of a POS filter to ignore invalid instances (see section 3.3) had a strong effect on recall, due to reduced noise in the activation of the graph. Recall significantly improved upon the modification of starting weights after analysis of static graph metrics, although precision fell slightly. This result (0.82) constitutes state of the art in graph-based WSD for biomedical text. The fall in precision was not unexpected, since the graph was no longer so

biased toward specific word senses. We also present a further experiment that incorporates a fall-back mechanism for test cases where the spreading activation did not produce a disambiguated output. This result (0.89) constitutes state of the art in overall unsupervised biomedical WSD. This allows our method to assign a single word-sense for every ambiguous word or surface-form. This fall-back alone achieves accuracy of 59%, comparing favourably with a default-sense approach (54.5%: McInnes et al, 2011).

Finally, by identifying bias in the graph toward specific senses in the test corpus, using an 80% subset of the MSH-WSD data set for training, and then modifying the rate of decay for problematic nodes, we achieved a significant boost to recall, and consequently to overall accuracy. We draw a distinction between this and other results since the testing was not blind, but was using the gold-standard corpus directly, to examine the portions of the graph that did not perform well in testing. We envisage that this may still be of practical use in real-world applications, by firstly developing an appropriate gold-standard, which in conjunction with analysis of the ontology instance graph, will result in optimal output.

The current results reflect the scope of spreading activation being set to the whole document. Only one sense of a word is recognised within that context, and documents containing multiple interpretations of the same word will not be correctly disambiguated. However, by configuring the scope to a sentence or paragraph we may reduce the potential accuracy of the output by decreasing the available instances for activation. Prior research into the "One sense per discourse" hypothesis suggests that the existing approach should be appropriate in up to 98% of cases (Gale et al, 1992).

Experiment Description	Precision	Recall	FMeasure	Accuracy
1. Baseline system	0.935	0.659	0.6639	0.62
2. Baseline + POS Filter	0.901	0.721	0.7872	0.74
3. As in 2, with SGA re-weight	0.841	0.822	0.8317	0.82
4. As in 3, confidence fallback	0.912	0.887	0.8995	0.89
5. SGA+WSD (20% test set)	0.986	0.942	0.9635	0.93
McInnes et al, 2011				0.72
J-Yepes & Aronson, 2012				0.87

Table 5. Comparison of WSD Results.

4.1 Identifying and Resolving Graph Bias or Imbalance

In experiment 5, having already identified specific cases that remained unbalanced, we attempted to rectify this by examining the graph in parallel with the WSD metric data. If a graph displays characteristics indicating imbalance or bias, for example where a node is unreachable (isolated in the graph), or node degree and node edge-type variation are relatively low (see section 3.5), it is less likely that the spreading activation will reflect the meaning of the text. We made discoveries similar to the following:

- 80% of nodes with Total ETV >15 had WSD precision of over 90%
- 60% of nodes with Total ETV <5 had precision of less than 10%

We also discovered cases in the graph where a node had very high Degree (> 100), and relatively low ETV. In terms of spreading activation, these nodes would be especially problematic. We have coined the term 'Black Hole Node' to describe this phenomenon. In psycholinguistic terms, this may be comparable to the notion of a Freudian slip, where a node in the graph which is not immediately relevant to the context of the document, has become over-stimulated by its connectivity, or as Norvig (1986) would suggest, its "promiscuity". The signal will gravitate towards such an over-connected node during the process of spreading activation, affecting the relevance of other nodes in that context. An example black hole node is the UMLS CUI C0035298, representing a retina in a human eye, with 1636 edges and 19 edge types. The extra noise in activating such a node can skew the signal across the entire graph. Word senses that compete for relevance with this or related nodes will have poorer accuracy. We modified the activation strategy to reflect this by increasing the rate of decay on such nodes from 0.5 to 0.99.

By ensuring that only the graph weighting strategy is modified, we can keep all word-senses present in the graph, resolving the issue identified by Norvig (1986) where such graph content had to be removed. Using the WSD metric output, we also modified the activation strategy to cope with bias toward particular senses in the test corpus. We reduced the starting weight for semantic categories for the high-scoring sense, in order to potentially increase the relative semantic importance of the alternative senses. Table 6 demonstrates some of the improvements achieved with regard to specific ambiguous terms.

Term	F-Measure Before	F-Measure After
Murine Sarcoma Virus	0	0.47
Gamma-Interferon	0.013	0.28
RA	0.021	0.59
CCD	0.033	1
AA	0.899	0.99

Table 6. Examples of term-specific improvement using re-weighting strategy.

4.2 MSH-WSD Data Set

In working with the MSH-WSD data set, we came across many issues that Navigli (2009) previously identified. The number of ambiguous senses (423) in the context of the full UMLS set of almost 3 million, reduces the validity of this corpus for measuring real-world viability and accuracy. Further to this, our results with lexical analysis optimisation demonstrate that the test corpus ignored surrounding context for potentially overlapping terms, such as “bat” and “fruit bat”. In such cases, it would have been more accurate to use the CUI for “fruit bat” as the specific type of “bat”, but the test corpus does not reflect this. Our algorithm is sensitive to contextual semantics, so ensuring that all lexical matches of any length remain present, potentially reduces the accuracy of the algorithm’s output, as well as the real-world utility of the approach. In spite of the various data transformation techniques applied, our recall maximised at 96.4%. Critically, when we normalize our overall accuracy (0.89) to take this into account, we reach accuracy of 0.92, a significant achievement in unsupervised WSD. We are currently examining what may be required to achieve maximum recall of 100%. While such a result is not guaranteed, without full coverage of the test set, we have not yet measured the full potential of this method.

4.3 Identifying Focus Areas for Ontology Improvement

One of the primary outcomes of this research is a method for the identification of specific ontology portions that require further development. As we have seen in section 4.1, there are several candidates which stand out. Other issues pointing to required enhancements in the ontology were around the notion of isolated nodes in the graph. An example of this is "ADA", the American Dental Association. It is surprising to discover that although this term’s associated CUI (C0002456) is listed in 7 source ontologies of the UMLS SN, there are no semantic relationships in the source between this CUI and any others. Of the 203 ambiguous terms in the MSH-WSD data set, 5 of those terms had associated nodes that were similarly isolated in the graph. Without any semantic relationship to other concepts, it is reasonable to suggest that the ontology would benefit from focused development of these nodes’ surrounding context.

In terms of the variation of connectivity, we quickly discovered using our simple graph metrics that the “SIB” or sibling relationship was extremely common. Consider the concept C0325089 representing the *felidae family* or the animal *cat*, which has 8 connections, but for which SIB is the only available link type. Hard-wiring siblings in this fashion, with no other link, is unhelpful since spreading activation can already identify siblings from common parent nodes. We contend that such concepts are not as well connected as they may first appear, and are therefore strong candidates for further development. This will not be apparent from the Degree metric alone, but by combining Degree and Edge Type Variation with node-specific accuracy in an NLP task, it becomes a straightforward process. Following this discovery, we also suggest that an empirical analysis of link quality would be beneficial, although this would not be a trivial task given the size of the data set (~3 million senses and ~700 link types).

5 Summary and Future Research

We have presented a new method for evaluating ontology semantics which has several advantages over existing approaches. We have shown how the application of graph theoretical analysis to semantic structures like ontologies is a valid means by which to assess their semantic quality, while enabling the recommendation of specific focus areas for further development. We have additionally demonstrated that a graph-metric based weighting strategy for spreading activation can overcome an ontology’s inherent semantic inconsistencies, facilitating the optimisation of the ontology for a given NLP task.

In the case of our UMLS prototype, we made significant improvements using this technique, achieving state of the art in unsupervised knowledge based WSD (0.82), as well as achieving state of the art in overall unsupervised WSD, with the use of a fall-back probability score (0.89). An additional semi-

supervised approach, leveraging gold-standard data from a training portion of the MSH-WSD data set, had very promising performance (0.93). The amount of required input data to this method is relatively small when compared with fully supervised approaches, as a single gold-standard annotation in each target context is sufficient to evaluate the graph using our spreading activation algorithm.

In future we would like to apply this technique to other ontologies, and associated test sets, for other domains in NLP. Merging of domain-specific ontologies with more general semantic resources like Yago or Wordnet may help to facilitate the activation of otherwise poorly connected or isolated nodes in the graph. We would like to investigate the automatic learning of an optimal spreading activation weighting strategy. An empirical study comparing data from human ontology reviewers with this spreading activation technique, would also be helpful.

We would like to expand the set of metrics used, by adapting other existing graph theoretical metrics to suit the requirements of NLP. Some promising examples are "Centrality" and "Betweenness" outlined by Brandes and Erlebach (2005), which determine the relative importance of a node within a graph. In the case of UMLS, we can perform a comprehensive static analysis of all ambiguous CUIs within the data set, identifying competing senses which do not have sufficient separation in the graph. These senses could then be targeted in the configuration of the spreading activation strategy.

As interest grows in the use of graph theoretical methods for the analysis of cognitive processes (Van Dijk et al, 2010; Bullmore and Sporns, 2009; Sporns, 2003), exploring the relationship between spreading activation in a graph representing ontology semantics, as performed in this research, and in neural activity during psycholinguistic experimentation (Fang and Evermann, 2010), becomes an exciting prospect that may lead to a better understanding of semantic processing in the human brain.

Acknowledgements

Sincere thanks to Mikhail Sogrin (IBM), the talented developer of both 'Galaxy' and the lexicon expansion framework used here, without whom this research would not have been possible.

References

Agirre, E., Martínez, D., de Lacalle, O. L., & Soroa, A. (2006, July). Two graph-based algorithms for state-of-the-art WSD. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (pp. 585-593). Association for Computational Linguistics.

Agirre, E., & Soroa, A. (2009, March). Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.* ACL

Agirre, E., Soroa, A., & Stevenson, M. (2010). Graph-based Word Sense Disambiguation of biomedical documents. *Bioinformatics*, *26*(22), 2889-2896.

Brandes, U., & Erlebach, T. (Eds.). (2005). Network analysis: methodological foundations (Vol. 3418). Springer.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience, 10 (3), 186-198.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. Journal of verbal learning and verbal behavior, 8(2), 240-247.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, *11*(6), 453-482.

Diestel, R. (2005), *Graph Theory* (3rd ed.), Berlin, New York: Springer-Verlag, ISBN 978-3-540-26183-4.

El-Rab, W. G., Zaïane, O. R., & El-Hajj, M. (2013, August). Biomedical text disambiguation using UMLS. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 943-947). ACM.

Evermann, J., & Fang, J. (2010). Evaluating ontologies: Towards a cognitive measure of quality. *Information Systems*, *35*(4), 391-403.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., & Welty, C. (2010). Building Watson: An overview of the DeepQA project. AI magazine, 31(3), 59-79.

Gale, W. A., Church, K. W., & Yarowsky, D. (1992, February). One sense per discourse. In Proceedings of the workshop on Speech and Natural Language (pp. 233-237). Association for Computational Linguistics.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2)

Guarino, N., Oberle, D., & Staab, S. (2009). What is an Ontology? In *Handbook on ontologies* (pp. 1-17). Springer

Jiménez-Ruiz, E., Grau, B. C., & Horrocks, I. (2012). Exploiting the UMLS Metathesaurus in the Ontology Alignment Evaluation Initiative. In *E-LKR Workshop* (pp. 1-6).

Jiménez Ruiz, E., Grau, B. C., Horrocks, I., & Berlanga, R. (2011). Supporting concurrent ontology development: Framework, algorithms and tool. *Data & Knowledge Engineering*, *70*(1), 146-164.

Jimeno Yepes, A., & Aronson, A. R. (2012, January). Knowledge-based and knowledge-lean methods combined in unsupervised word sense disambiguation. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* (pp. 733-736). ACM.

Liang, T., & Lin, Y. H. (2005). Anaphora resolution for biomedical literature by exploiting multiple resources. In Natural Language Processing–IJCNLP 2005(pp. 742-753). Springer Berlin Heidelberg

Ma, Y., Jin, B., Liu, X., Liu, L., & Lu, K. (2013). A Graph Derivation Based Approach for Measuring and Comparing Structural Semantics of Ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 1.

McInnes, B. T., Pedersen, T., Liu, Y., Melton, G. B., & Pakhomov, S. V. (2011). Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity. In *AMIA Annual Symposium Proceedings* (Vol. 2011, p. 895). American Medical Informatics Association.

McInnes, B. T., & Pedersen, T. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. Journal of biomedical informatics, 46(6), 1116-1124.

National Library of Medicine. 2013. *Unified Medical Language System*, version 2013AA. NLM

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, *41*(2), 10.

Norvig, P. (1986). Unified theory of inference for text understanding. CALIFORNIA UNIV BERKELEY GRADUATE DIV.

Noy, N. F. (2004). Tools for mapping and merging ontologies. In *Handbook on ontologies* (pp. 365-384). Springer

Pace-Sigge, M. (2013). Lexical Priming in Spoken English Usage. Palgrave Macmillan.

Pisanelli, D. M., Gangemi, A., & Steve, G. (1998). An ontological analysis of the UMLS Metathesaurus. In *Proceedings of the AMIA symposium* (p. 810). American Medical Informatics Association.

Plaza, L., Jimeno-Yepes, A. J., Díaz, A., & Aronson, A. R. (2011). Studying the correlation between different word sense disambiguation methods and summarization effectiveness in biomedical texts. *BMC bioinformatics*

Quillian, M.R. (1966). Semantic Memory. Unpublished doctoral dissertation, Carnegie Institute of Technology (Re-printed in part in M. Minsky (1968). Semantic Information Processing. Cambridge, Mass. MIT Press).

Sicilia, M. A., Rodríguez, D., García-Barriocanal, E., & Sánchez-Alonso, S. (2012). Empirical findings on ontology metrics. *Expert Systems with Applications*, *39*(8), 6706-6711.

Sporns, O. (2003). Graph theory methods for the analysis of neural connectivity patterns. In *Neuroscience Databases* (pp. 171-185). Springer US.

Tartir, S., Arpinar, I. B., Moore, M., Sheth, A. P., & Aleman-Meza, B. (2005, November). OntoQA: Metric-based ontology quality analysis. In *IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources* (Vol. 9).

Troussov, A., Sogrin, M., Judge, J., & Botvich, D. (2008). Mining socio-semantic networks using spreading activation technique. In *Proc. International Workshop on Knowledge Acquisition from the Social Web*

Tsatsaronis, G., Vazirgiannis, M., & Androutsopoulos, I. (2007, January). Word Sense Disambiguation with Spreading Activation Networks Generated from Thesauri. In *IJCAI* (Vol. 7, pp. 1725-1730).

Van Dijk, K. R., Hedden, T., Venkataraman, A., Evans, K. C., Lazar, S. W., & Buckner, R. L. (2010). Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. Journal of neurophysiology, 103(1), 297.

Vrandečić, D., & Sure, Y. (2007). How to design better ontology metrics. In *The Semantic Web: Research and Applications* (pp. 311-325). Springer Berlin Heidelberg.