Authors accepted manuscript

Ontology network analysis for safety learning in the railway domain

Miguel Figueres-Esteban, Peter Hughes, Coen van Gulijk

Institute of Railway Research, University of Huddersfield, Huddersfield, UK

ABSTRACT: Ontologies have been used in diverse areas such as Knowledge Management (KM), Artificial Intelligence (AI), Natural Language Processing (NLP) and Semantic Web as they allow software applications to integrate, query and reason about concepts and relations within a knowledge domain. For Big Data Risk Analysis (BDRA) in railways, ontologies are a key enabler for obtaining valuable insights into safety from the large amount of data available from the railway. Traditionally, the ontology building has been an entirely manual process that has required a considerable human effort and development time. During the last decade, the information explosion due to the Internet and the need to develop large-scale methods to extract patterns in a systematic way, has given rise the research area of "ontology learning". Despite recent research efforts, ontology learning systems are still struggling with extracting terms (words or multiple-words) from text-based data. This manuscript explores the benefits of visual analytics to support the construction of ontologies for a particular part of railway safety management: possessions. In railways, possession operations are the protection arrangements for engineering work that ensure track workers remain separated from moving trains. A network of terms from possession operations standards is represented to extract the concepts of the ontology that enable the safety learning from events related to possession operations.

1 INTRODUCTION

1.1 Knowledge management for safety

The Big Data Risk Analysis research programme is a joint effort between the University of Huddersfield and RSSB that aims to support safety decision-making for risk analysis using data-analytical techniques. This paper focuses on semi-automated extraction of relevant organisational knowledge based on Visual Analytics, in particular, capturing knowledge for track possession safety. In railways, possession operations are the protection arrangements for engineering work that ensure track workers remain separated from moving trains. Possession management is particularly relevant for this study since it is a safety critical operation - workers' lives are at stake - but there is a large body of documentation within the railway from which to extract information.

1.2 The need to capture knowledge

Railway is a complex system due to the large number of interacting subsystems. Each subsystem can be managed simultaneously by different organisations (e.g. infrastructure managers, operators, manufacturers or maintainers). Furthermore, each organisation can have its own structure (e.g. financial area, health and safety area, resources area or selling area) formed by people of different expertise, skills and competences that change over time and don't necessarily draw from the same jargon. For example, a passenger might be a ticket buyer for an economist but equally, it might be an individual on a moving

train for a safety expert (even if they didn't pay their ticket). Thus, a railway is a rich tapestry of different organisational knowledge, created for different purposes and people in many different contexts.

For that reason it is necessary to organise and manage the heterogeneous knowledge to obtain insight from the large amount data available in different information systems. The most common technique used by computer scientists to represent a common framework of understanding and manage the knowledge is an ontology.

1.3 What is an ontology

A formal, and broadly accepted, definition of an ontology is provided by Gruber (1995):

"An ontology is an explicit specification of a conceptualization."

Guarino et al. (1997) classify ontologies based on the structure and subject of the conceptualization. Different categories of ontologies such as domain ontologies or application ontologies can be found depending on the reusability or specificity of the knowledge.

Ontologies have been used in diverse areas such as Knowledge Management (KM), Artificial Intelligence (AI), Natural Language Processing (NLP) and Semantic Web as they allow software applications to integrate, query and reason about concepts and relations within a knowledge domain. For instance, Bloehdorn & Hotho (2004) demonstrate statistically significant improvements on a text classification task with the use of an ontology. In the railway domain, the FP6 European Integrail project (http://www.integrail.eu/) and the RailML community (http://www.railml.org) proved the utility of ontologies in the communication and integration of data through railway information systems (Van Gulijk & Figueres-Esteban 2016).

For BDRA, ontologies are a key enabler not only to perform risk analysis, but also for search engine processes (queries), data integration from heterogeneous databases and analysis of text in railway safety documentation (Van Gulijk et al. 2015).

1.4 Ontology learning

Ontology building has been traditionally a manual process that requires considerable effort from knowledge engineers and domain experts. During the last decade, the information explosion due to the Internet and the need to develop large-scale methods to extract patterns in a systematic way, has given rise the research area of "ontology learning" (Maedche & Staab 2001).

Figure 1 provides an overview of the steps for building an ontology in the BDRA program. Each stage of the process creates an ontology at a different level of abstraction, encoding knowledge of different expressiveness and specificity. The first steps create lightweight ontologies that define the main concepts of an ontology (e.g. extracting terms and building glossaries and thesauruses) and their taxonomical relationships (e.g. building UML class diagrams). The last steps are focused on heavyweight ontologies that represent non-taxonomical relationships using formal languages and ontology languages (e.g. RDF/RDF's and OWL-DL).

The "blocks" or "atoms of knowledge" of the ontology come from the first stage: the term extraction. The aim of this stage is to represent a common framework of understanding of a specific domain (vocabulary). There are two ways to define the terms (words or multi-words) of an ontology. The first is by means of

expert's elicitation, where domain experts decide the best terms to include in the ontology (Uschold et al. 1998). The second extracts terms from the organisational knowledge represented in text documents (e.g. specifications or standards). The term extraction is based on linguistic (e.g. part-of-speech tagging) or statistical approaches based on frequency and co-occurrence affinity (Biemann 2005; Buitelaar et al. 2005; Wong et al. 2012). The final result is a lightweight ontology in the form of list of terms that represent the key concepts of a domain.

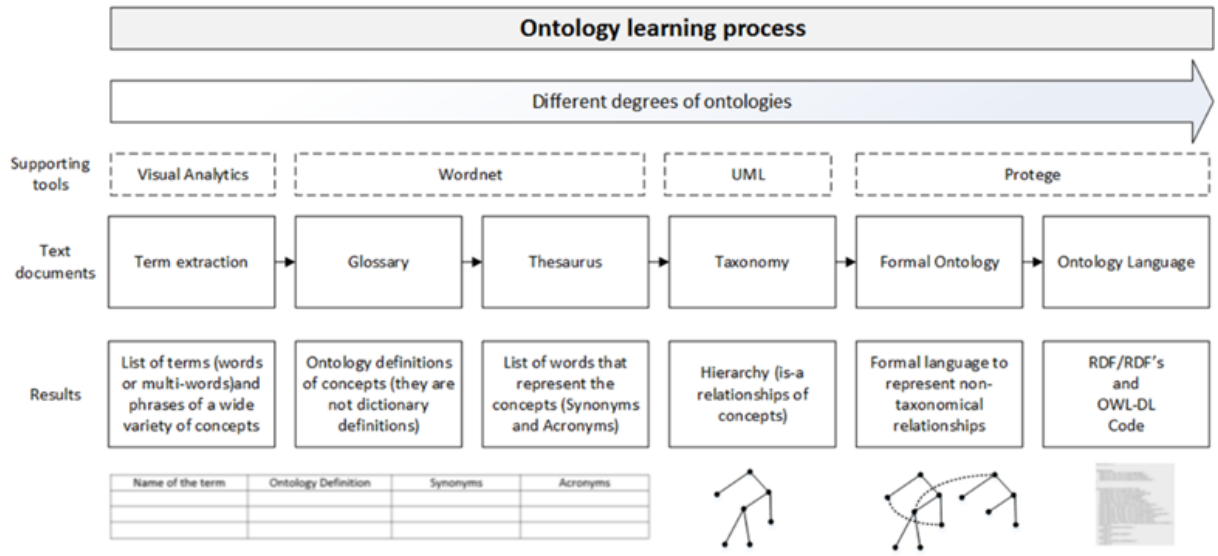

Figure 1. Ontology leaning process.

1.5 Visual analytics

Visual Analytics (VA) is the discipline that combines "…automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets." (Keim et al. 2008). VA can be interpreted as a variant of 'big data analytics' supported by interactive visualisation techniques, that is, VA not only displays results but also drives the analysis process through human judgment (Figueres-Esteban 2015; Figueres-Esteban et al. 2015a; Thomas & Cook 2005).

This manuscript is focused on how VA can support the term extraction process for building ontologies. It is beyond the scope of this manuscript to describe the complete ontology learning process, rather this paper provides the context to explore VA techniques.

1.6 Text analysis

Three different computing text analysis approaches have been identified for retrieving information from text: thematic, semantic and networks (Popping 2000). Thematic analysis has been the main approach for a long time and it is based on the frequency of concepts (e.g. words or "bag of words") that allows classification of topics of texts. Semantic analysis takes into account the relationships among the concepts encoding the semantic grammar (e.g. subject, verb and object). Network analysis is based on network text

analysis to obtain semantically linked concepts. VA techniques supports the third approach to improve the interpretation of text (Figueres-Esteban et al. 2015b; Paranyushkin 2011).

This manuscript explores the benefits of network text analysis to support the construction of ontologies in relation to railway track possession management. A network of terms from a possession management standard is represented to extract the terms of the ontology that will enable the safety learning from events related to possession operations.

2 METHODOLOGY

The railway Rule Book GE/RT8000/T3 (Possession of a running line for engineering work) was selected as information source to extract terms related possession operations. This document is in PDF format with a particular and complex style layout (different types of heads, embedded graphs and margin notes). Although a Python PDF parser was used to extract automatically the text, there was nevertheless need for further manual cleansing of the text.

The NLTK and NetworkX toolkits (Bird et al. 2009; Hagberg et al. 2015) were used to pre-process the text and create the word network. The text was tokenised and tagged using the Stanford Tokeniser and Part-Of-Speech Tagger (Hughes et al. 2015). A context window of size two and distance weighting equal to one was selected to create the directed network (Paranyushkin 2011; Sahlgren 2006). Using these settings the network considered only the links between adjacent words in the right direction and the value of each link was equal to one. Each node corresponds to a word with the Part-Of-Speech tag attribute (Penn Treebank tag).

Gephi software was used to visualise the network of words and perform the analysis based on the degree of the nodes (number of links connecting a node), the weight of the links (the frequency of links between two nodes) and the Part-Of-Speech tag attribute.

The main candidates to represent concepts of an ontology are the nouns or combination of them. The network was filtered using the regular expression that represents the lexico-syntactic pattern of a combination of nouns (tagged with the NN and NNS labels) or proper nouns (tagged with the NNP or NNPS labels).

3 RESULTS

The resulting network is a directed graph of 542 nodes (words) and 1947 links. The syntactic pattern analysis provided a set of clusters. Figure 2 shows some of those clusters. The first cluster represents a large conglomerate of isolated nodes (Figure 2.a). The remaining clusters represent small graphs that provide multi-word relationships such as bigrams (Figure 2.b, 2.c, and 2.c).

The cluster of isolated nodes shows high degree nodes such as ‘PICOP’, ‘signaller’, ‘points’, ‘WSMB’, ‘SWL’, ‘ES’, ‘instructions’ or ‘section’.

The cluster of bigrams (Figure 2.b) shows strong relationships between the nodes {‘Train’, ‘Register’}, {‘Operations’, ‘Control’}, {‘road’, ‘traffic’} and {‘driver’, ‘permission’} and lower relationships between the

nodes {‘axle’, ‘counter’}, {‘maximum’, ‘speed’} and {‘ENGINEERING’, ‘SUPERVISOR’}. The highest degree nodes are driver and permission, although all nodes have a similar degree.

The cluster of the Figure 2.c has strong links between {‘work’, ’site’}, {‘Engineering’, ‘Notice’}, {‘Weekly’, ‘Operating’} and {‘Operating’, ‘Notice’}. The higher degree nodes are ‘line’, ‘work’, ‘site’, ‘engineering’ and ‘trains’.

The cluster of the Figure 2.d shows strong relationships between the nodes {‘block’, ‘marker’}, {‘detonator’, ‘protection’}, {‘possession’, ‘arrangements’) and {‘movement’, ‘authority’}. The highest degree nodes are ‘movement’, ‘possession’, ‘protection’, ‘detonator’, ‘block’ and ‘marker’.

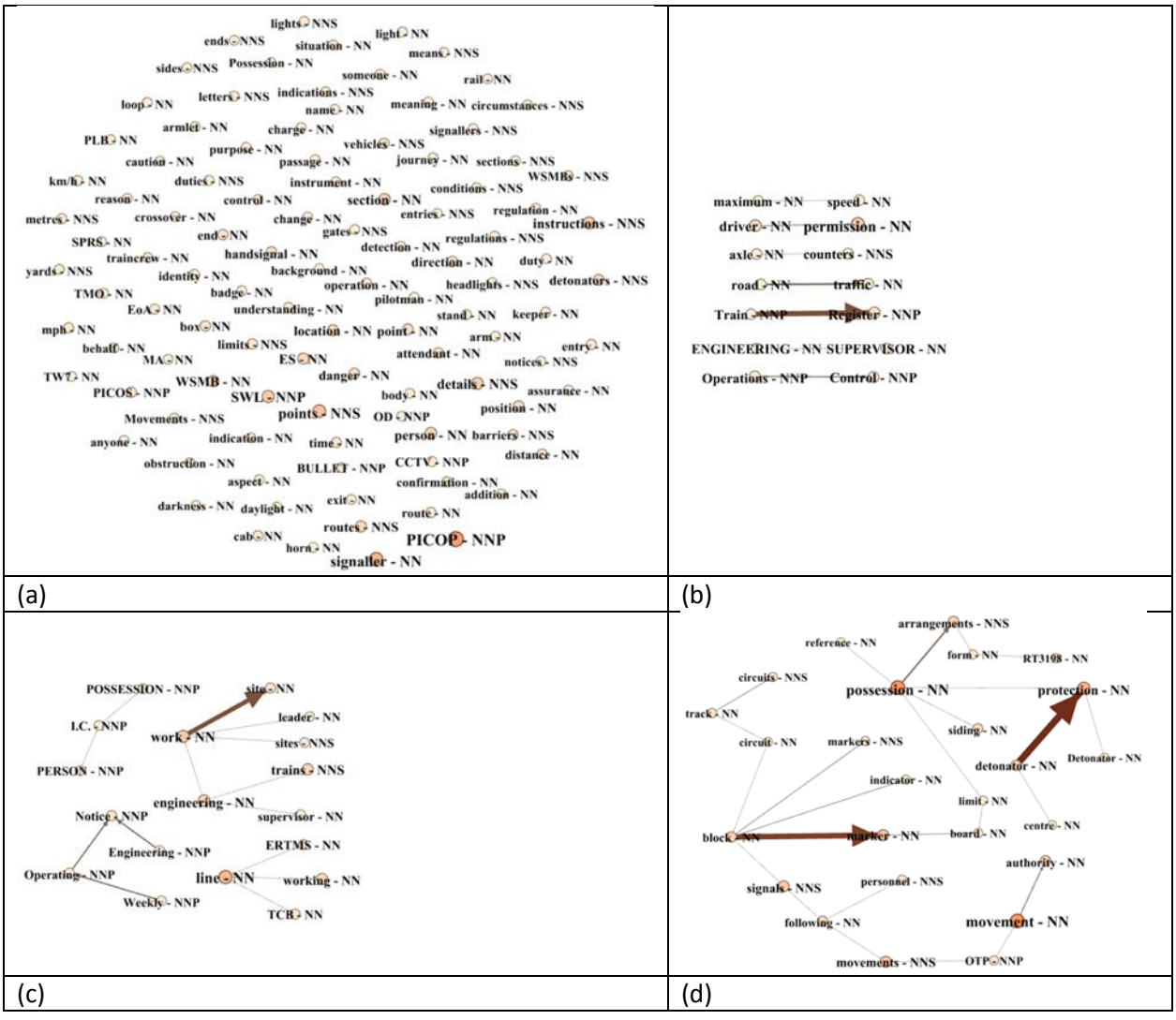

Figure 2. Sample of clusters that represent a set of nodes (words) and the relationships between them. The size of each node represents its degree.

4 DISCUSSION

The lexico-syntactic analysis displays a set of clusters formed by nodes that represent words from the possession standard. Some of these nodes represent the same concept such as {‘light’, ‘lights’}, {‘detonator’, ‘Detonator’, ‘detonators’}, {‘Engineering’, ‘engineering’}, {‘circuit’, ‘circuits’} or {‘signaller’, ‘signallers’}. This results occurs since the pre-processing did not include the conversion into a common case (for example lowercases), nor text stemming. The network facilitates the detection of acronyms (e.g. ‘PICOP’, ‘SWL’ or ‘ES’) and facilitates the building of paths that provide important concepts represented by multi words such {‘Weekly’→ ‘Operating’→ ‘Notice’}, {‘PERSON’ → ‘I.C.’ → ‘POSSESSION’} or {‘RT3198’ → ‘form’}. Moreover, the network allows detection of different ways to express the same concept such as {‘PICOP’, ‘PERSON’ → ‘I.C.’ → ‘POSSESSION’}. The drawback of this approach is that the weight of some paths is dispersed. In some cases, strong relationships between nodes can provide a clear indication of a candidate concept, for example {'block', 'marker'}. This strong relationship can help identify other, weaker, relationships such as {'block', 'markers'}. However an advantage of VA is that it allows concepts to be identified even when only weak relationships exist; for example {'track', 'circuit'} and {'track', 'circuit's} are both weak relationships, yet when viewed together they indicate a candidate concept. Identifying a concept in this way based only on weak relationships is a clear advantage of VA.

4.1 Cluster description

The resulting clusters provide different types of information from the ontology learning perspective. The cluster of isolated nodes provides a list of words that represent important concepts. The higher degree nodes of this cluster essentially show concepts related to people such as ‘PICOP’ (person in charge of possession), ‘signaller’, ‘SWL’ (safe work leader) or ‘ES’ (engineering supervisor) and related to places such as ‘points’ or ‘section’. The high degree indicates the importance of the concept in the possession domain and indicated that they are closely related to many other different concepts, and therefore, they appear in different types of contexts (e.g. different sections of the standard). Moreover, this cluster pro-vides a good list of words that describe the same concepts, which support the construction of a lexicon that will be useful in the next steps of the ontology building.

The cluster represented in the Figure 2.b shows a very good candidate of concepts represented by multi-words such as ‘Train Register’, ‘Operations Control’, ‘road traffic’, ‘driver permission’, ‘axle counter’ and ‘Engineering Supervisor’. The cluster allows to improve the tokenisation process described in Hughes et al. (2015) and it enables future iterations to detect new types of multi-words.

The last two clusters (Figure 2.c and 2.d) represent isolated pieces of the network. These clusters need additional interpretation since it is necessary to analyse different paths to detect multi-words greater than two. The Gephi interactive environment provides a view of the paths of one node, where it is possible to see multi-word terms such as ‘Weekly Operating Notice’, ‘work site’, ‘Engineering Notice’, ‘block marker’, ‘detonator protection’, ‘siding possession’, ‘track circuit’ and ‘possession reference’. Again, the process supports the creation of a vocabulary related to the possession domain and provides valuable information to improve the tokenisation process.

4.2 VA supporting term extraction.

Despite recent research efforts, ontology learning systems are still struggling with extracting terms (words or multiple-words), taxonomic relationships (e.g. hyponym or synonym) and non-taxonomic relationship's in arbitrary knowledge domains (Brewster et al. 2002). Statistical approaches are strongly dependent on the frequency of terms in large corpora of text, which can mean that low-frequency terms are ignored regardless of their relevance to the domain. Unfortunately, low frequency multi-words tend to form a significant proportion of the total terms in most texts (Piao et al. 2005).

The approach in this work is different in two ways. Firstly, it explores the benefits of visual analytics to support term extraction when traditional statistical methods would not be suitable since a large corpora of text of the domain is not available. The interactive visualisation environment allows analysts detection of even low-weighted relationships such as 'track circuit' or 'limit board'. This interaction reduces the scattering effect produced by the pre-processing as explained above.

Secondly, the approach reduces the effort required by safety analysts and increases the accuracy to detect important terms from text of a specific domain such as railway, where the identification and use of key concepts and relationships is more critical than if it were an arbitrary knowledge domain.

4.3 VA supporting human interaction

VA is the combination of a set of disciplines that enable data analysis through the interaction of visual representations with human judgment (Figueres-Esteban et al. 2015a). In this case study, the Gephi software has been used to represent the networks of words and interact with them to detect important concepts that would be part of more complex ontologies. However, this is an iterative process. This manuscript has shown a first list of candidates of concepts and multi-word expressions. This list of words would be a source of new tokens for the tokenisation process. Therefore, new networks of words would have to be represented in order to detect new types of concepts or multi-words expressions, refining progressively the list of terms that represent the possession ontology. The criteria to validate and decide the right number of iterations is beyond the scope of this paper since it remains in the ontology field.

4.4 Limitations and future work

Although the open-source VA software allows for text analysis, specific interactions between the graph and the text pre-processing are not possible. For example, interactive functions such as joining two nodes to create new tokens from the graph or automatically updating the network with new tokens would speed up the visual analysis. New specific VA tools for text analysis should be developed for supporting the ontology learning process.

From the ontology perspective, the obtained lightweight ontology comes from a formal document that describes processes related to possession operations. It provides a limited lexicon of concepts that could not match with information on the same topic but from different sources (e.g. Close Call system, Weekly Operating Notices or accident reports related to possession operations). The current results would be a good starting point in order to extend and complete the possession ontology from alternative text data sources that could provide new concepts and enrich the lexicon (e.g. new synonyms, abbreviations and informal or slang terms).

5 CONCLUSIONS

This paper discusses a new research framework based on VA techniques for railway safety domain. In particular, to support the ontology learning process from unstructured text data. The results demonstrate that mature network analysis techniques for text analysis can be applied to extract safety-relevant terms that will be the future 'blocks' of a formal ontology for risk analysis.

On one hand, this method needs additional interpretation of safety analyst to detect the right terms related to the domain. On the other hand, it improves considerably the efficiency of the term extraction and avoid the low frequency gap of the statistical approaches.

This work also demonstrates new possibilities of improving the NLP techniques, such as the creation of new tokens, which can improve the text analysis techniques for risk analysis.

6 ACKNOLEDGEMENTS

The work reported in this paper was undertaken under the Strategic Partnership between the University of Huddersfield and RSSB.

7 REFERENCES

Biemann, C., 2005. Ontology Learning from Text: A Survey of Methods. LDV-Forum, 20, pp.75–93.

Bird, S., Klein, E. & Loper, E., 2009. Natural Language Processing with Python, O'Reilly Media, Inc.

Bloehdorn, S. & Hotho, a., 2004. Text classification by boosting weak learners based on terms and concepts. Fourth IEEE International Conference on Data Mining (ICDM'04), pp.4–7.

Brewster, C., Ciravegna, F. & Wilks, Y., 2002. User-Centred Ontology Learning for Knowledge Management. In LNCS 2553. pp. 203–207.

Buitelaar, P., Cimiano, P. & Magnini, B., 2005. Ontology Learning from Text : An Overview. In Ontology Learning from Text: Methods, Applications and Evaluation. IOS Press, pp. 3–12.

Figueres-Esteban, M., 2015. Visualisation and Risk Communication in Railway Big Data Risk Analysis (BDRA): Literature Review.

Figueres-Esteban, M., Hughes, P. & Van Gulijk, C., 2015a. The role of data visualization in Railway Big Data Risk Analysis. In Proceedings of the 25th European Safety and Reliability Conference, ESREL 2015. p. 7.

Figueres-Esteban, M., Hughes, P. & Van Gulijk, C., 2015b. Visualising Close Call in railways: a step towards Big Data Risk Analysis. In Fifth International Rail Human Factors Conference. pp. 725–734.

Gruber, T.R., 1995. Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, 43, pp.907–928. Available at: http://dx.doi.org/10.1006/ijhc.1995.1081.

Guarino, N., 1997. Understanding, building and using ontologies. International Journal of Human Computer Studies, 46, pp.293–310.

Van Gulijk, C. et al., 2015. Big Data Risk Analysis for Rail Safety? In Proceedings of the 25th European Safety and Reliability Conference, ESREL 2015.

Van Gulijk, C. & Figueres-Esteban, M., 2016. Background of Ontology for BDRA.

Hagberg, A., Schult, D. & Swart, P., 2015. NetworkX Reference.

Hughes, P., Van Gulijk, C. & Figueres-Esteban, M., 2015. Learning from text-based close call data. In Proceedings of the 25th European Safety and Reliability Conference, ESREL 2015. p. 8.

Keim, D. et al., 2008. Visual analytics: Definition, process, and challenges. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 154–175.

Maedche, A. & Staab, S., 2001. Ontology Learning for the Semantic Web. IEEE Intelligent Systems, pp.72–79.

Paranyushkin, D., 2011. Identifying the Pathways for Meaning Circulation using Text Network Analysis. Nodus Labs, p.26.

Piao, S.S. et al., 2005. Comparing and combining a semantic tagger and a statistical tool for MWE extraction. Computer Speech and Language, 19, pp.378–397.

Sahlgren, M., 2006. Towards pertinent evaluation methodologies for word-space models. In Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC. pp. 821–824.

Thomas, J.J. & Cook, K.A., 2005. Illuminating the path: The research and development agenda for visual analytics, IEEE Computer Society.

Uschold, M. et al., 1998. The enterprise ontology. The knowledge engineering review, 13, pp.31–89.

Wong, W., Liu, W. & Bennamoun, M., 2012. Ontology learning from text. ACM Computing Surveys, 44, pp.1–36.