From DBpedia and WordNet hierarchies to LinkedIn and Twitter

Aonghus McGovern
ADAPT Centre
Trinity College Dublin
Dublin, Ireland
amcgover@scss.tcd.ie

Alexander O'Connor
ADAPT Centre
Trinity College Dublin
Dublin, Ireland
Alex.OConnor@scss.tcd.ie

Vincent Wade
ADAPT Centre
Trinity College Dublin
Dublin, Ireland
Vincent.Wade@cs.tcd.ie

Abstract

Previous research has demonstrated the benefits of using linguistic resources to analyze a user's social media profiles in order to learn information about that user. However, numerous linguistic resources exist, raising the question of choosing the appropriate resource. This paper compares Extended WordNet Domains with DBpedia. The comparison takes the form of an investigation of the relationship between users' descriptions of their knowledge and background on LinkedIn with their description of the same characteristics on Twitter. The analysis applied in this study consists of four parts. First, information a user has shared on each service is mined for keywords. These keywords are then linked with terms in DBpedia/Extended WordNet Domains. These terms are ranked in order to generate separate representations of the user's interests and knowledge for LinkedIn and Twitter. Finally, the relationship between these separate representations is examined. In a user study with eight participants, the performance of this analysis using DBpedia is compared with the performance of this analysis using Extended WordNet Domains. The best results were obtained when DBpedia was used.

1 Introduction

Natural Language Processing (NLP) techniques have been shown in studies such as Gao et al. (2012) and Vosecky et al. (2013) to be successful in extracting information from a user's social media profiles. In these studies, linguistic resources are employed in order to perform NLP tasks such as identifying concepts, Named Entity Recognition etc. However, as Chiarcos et al. argue, significant interoperability issues are posed by the fact that linguistic resources 'are not only growing in number, but also in their heterogeneity' (2014). They further argue that the best way to overcome these issues is by using linguistic resources that conform to Linked Open Data principles. Chiarcos et al. divide such resources into two categories, distinguishing between strictly lexical resources such as WordNet and general knowledge bases such as DBpedia.

The study described in this paper examines a single resource of each type. WordNet is considered to be a representative example of a purely lexical resource given the extent of its use in research[1]. DBpedia is considered to be a representative example of a knowledge base because its information is derived from Wikipedia, the quality of whose knowledge almost matches that of Encyclopedia Britannica (Giles, 2005). These resources are compared by means of an investigation of the relationship between users' descriptions of their knowledge and background on LinkedIn with their description of the same characteristics on Twitter.

Both LinkedIn and Twitter allow users to describe their interests and knowledge by: (i) Filling in profile information (ii) Posting status updates. However, the percentage of users who post status updates on LinkedIn is significantly lower than the percentage of users who do so on Twitter (Bullas, 2015). On the other hand, LinkedIn users fill in far more of their profiles on average than Twitter users (Abel, Henze, Herder, and Krause, 2010).

[1] A list of publications involving WordNet: http://lit.csci.unt.edu/~wordnet/

Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015), pages 1–10,
Beijing, China, July 31, 2015.

Given the different ways in which users use these services, it is possible that they provide different representations of their interests and knowledge on each one. For example, a user may indicate a new-found interest in 'Linguistics' through their tweets before they list this subject on their LinkedIn profile. This study examines the relationship between users' descriptions of their interests and knowledge on each service.

2 Related Work

Hauff and Houben describe a study that investigates whether a user's bookmarking profile on the sites Bibsonomy, CiteULike and LibraryThing can be inferred using information obtained from that user's tweets (2011) . Hauff and Houben generate separate 'knowledge profiles' for Twitter and for the bookmarking sites. These profiles consist of a weighted list of terms that appear in the user's tweets and bookmarking profiles, respectively. The authors' approach is hindered by noise introduced by tweets that are unrelated to the user's learning activities. This problem could be addressed by enriching information found in a user's profiles with structured data in a linguistic resource.

However, there are often multiple possible interpretations for a term. For example, the word 'bank' has entirely different interpretations when it appears to the right of the word 'river' than when it appears to the right of the word 'merchant'. When linking a word with information contained in a linguistic resource, the correct interpretation of the word must be chosen. The NLP technique Word Sense Disambiguation (WSD) addresses this issue. Two different approaches to WSD are described below.

Magnini et al. perform WSD using WordNet as well as domain labels provided by the WordNet Domains project[2] (2002). This project assigned domain labels to WordNet synsets in accordance with the Dewey Decimal Classification. However, WordNet has been updated with new synsets since Magnini et al.'s study. Therefore, in the study described in this paper, the Extended WordNet Domain labels created by González et al. (2012) are used. Not only do these labels provide greater synset coverage, González et al. report better WSD performance with Extended WordNet Domains than with the original WordNet Domains.

Mihalcea and Csomai describe an approach for identifying the relevant Wikipedia articles for a piece of text (2007). Their approach employs a combination of the Lesk algorithm and Naïve Bayes classification. Since DBpedia URIs are created using Wikipedia article titles, the above approach can also be used to identify DBpedia entities in text.

Magnini et al's approach offers a means of analysing the skill, interest and course lists on a user's LinkedIn profile with regard to both WordNet and DBpedia. Unambiguous items in these lists can be linked directly with a WordNet synset or DBpedia URI. These unambiguous items can then provide a basis for interpreting ambiguous terms. For example, the unambiguous 'XML' could be linked with the 'Computer Science' domain, providing a basis for interpreting the ambiguous 'Java'.

The above analysis allows for items in a user's tweets and LinkedIn profile to be linked with entities in WordNet/DBpedia. Labels associated with these entities can then be collected to form separate term-list representations for a user's tweets and LinkedIn profile information.

Plumbaum et al. describe a Social Web User Model as consisting of the following attributes: Personal Characteristics, Interests, Knowledge and Behavior, Needs and Goals and Context (2011). Inferring 'Personal Characteristics' (i.e. demographic information) from either a user's tweets or their LinkedIn profile information would require a very different kind of analysis from that described in this paper, for example that performed by Schler and Koppel (2006). As Plumbaum et al. define 'Behaviour' and 'Needs and Goals' as system-specific characteristics, information about a specific system would be

[2] http://wndomains.fbk.eu/

required to infer them. ‘Context’ requires information such as the user’s role in their social network to be inferred.

Given the facts stated in the previous paragraph, the term lists generated by the analysis in this study are taken as not describing ‘Personal Characteristics’, ‘Behavior’, ‘Needs and Goals’ and ‘Context’. However, a term-list format has been used to represent interests and knowledge by Ma, Zeng, Ren, and Zhong (2011) and Hauff and Houben (2011), respectively. As mentioned previously, users’ Twitter and LinkedIn profiles can contain information about both characteristics. Thus, the term lists generated in this study are taken as representing a combination of the user’s interests and knowledge.

3 Research Questions

3.1 Research Question 1

This question investigates the possibility that a user may represent their interests and knowledge differently on different Social Media services. It is as follows:

RQ1. *To what extent does a user’s description of their interests and knowledge through their tweets correspond with their description of the same characteristics through their profile information on LinkedIn?*

For example, 'Linguistics’ may be the most discussed item in the user’s LinkedIn profile, but only the third most discussed item in their tweets.

This question is similar to that investigated by Hauff and Houben (2011). However, there is an important difference. This question does not try to determine whether a user’s LinkedIn profile can be inferred from their tweets. Instead, it investigates the extent of the difference between the information users give through each service.

3.2 Research Question 2

Studies such as Abel et al. (2011; 2012) show that information found in a user’s tweets can be used to recommend items to them e.g. news articles. Furthermore, as user activity is significantly higher on Twitter than on LinkedIn (Bullas, 2015), users may discuss recent interests on the former without updating the latter. The second research question of this study is as follows:

RQ2. *Can information obtained from a user’s tweets be used to recommend items for that user’s LinkedIn page?*

These questions aim to investigate: (i) The variation between a user’s description of their interests and knowledge through their LinkedIn profile and their description of these characteristics through their tweets (ii) Whether a user’s tweets can be used to augment the information in their LinkedIn profile.

4 Method

The user study method is applied in this research. This decision is taken with reference to work such as Lee and Brusilovsky (2009) and Reinecke and Bernstein (2009). The aforementioned authors employ user studies in order to determine the accuracy with which their systems infer information about users.

4.1 Analysis

The analysis adopted in this study consists of four stages:

- Identify keywords
- Link these keywords to labels in DBpedia /Extended WordNet Domains)
- Generate separate representations of the user’s interests and knowledge for their tweets and LinkedIn profile information
- Examine the relationship between these separate representations

4.2 Keyword Identification

The user’s LinkedIn lists of skills, interests and courses are treated as lists of keywords with respect to each resource. However, the process for identifying keywords in text (e.g. a textual

description on LinkedIn, a tweet) differs for each resource. For the DBpedia approach potential keywords derive from a precompiled list i.e. the list of all Wikipedia keyphrases. In the case of Extended WordNet Domains, no such list exists, meaning a different approach must be used for identifying keywords. Ellen Riloff describes various methods for identifying important items in a text (1999). Riloff describes how case frames can be used to identify nouns describing entities such as perpetrators and victims of crimes. A case frame approach cannot be adopted here as it relies on previous knowledge of the text being processed. Instead, each text is parsed in order to extract its noun phrases. These noun phrases are then investigated using n-grams for keywords that can be linked with WordNet synsets.

Keywords relating to Named Entities of the following types are ignored: Person; Place; Organization; Event; Animal; Film; Television Show; Book; Play (Theatre). This is because in a list of top LinkedIn skills compiled by *LinkedIn Profile Services*[3], not a single Named Entity of these types appears.

Any links appearing in text are extracted and cleaned of HTML. The remaining text is analysed in the manner detailed in the two previous paragraphs.

4.3 Linking keywords with labels

Ma et al. discuss methods for identifying user interests by combining information from different sources, including LinkedIn and Twitter (2011). The authors argue that, with the help of domain ontologies, texts a user has written can be used to identify both explicit and implicit interests. Explicit interests are identified by linking text items with ontology classes. Implicit interests are then identified by obtaining the parent and/or child of the identified ontology class. For example, consider an ontology in which 'Knowledge Representation' is the parent of 'Semantic Web'. If a user explicitly indicates they are interested in 'Semantic Web', they are implicitly indicating that they are interested in 'Knowledge Representation'. However, Ma et al. provide the caveat that only the immediate parents and children of a particular class (i.e. one level above/below) should be identified for a particular class.

In the study described in this paper, explicit information is obtained by linking keywords with labels in DBpedia/Extended WordNet Domains. Unambiguous keywords are linked directly. Ambiguous keywords are linked to labels using the methods described in the 'Related Work' section. Implicit information is obtained by identifying parent class(es) only. This decision was taken with reference to the 'is-a' subsumption relation. Under this relation, if an object B inherits from an object A, all instances of B are instances of A. However, instances of A are not necessarily instances of B. For example, if a user explicitly expresses an interest in 'Knowledge Representation' they are not necessarily implicitly expressing an interest in 'Semantic Web'.

4.4 Representation of user interests and knowledge

User interests and knowledge are represented as weighted lists of terms. Weighting schemes such as tf-idf are not used because as Hauff and Houben argue, such measures are not best suited to measuring the relative importance of terms for a user. The authors describe the inherent problem with measures such as tf-idf: 'if a tenth of the CiteULike articles in our index for example would include the term genetics, it would receive a low weight, although it may actually represent the user's knowledge profile very well' (2011). The procedure for calculating term weights in this study is thus identical to that in Hauff and Houben's study. A term's weight is calculated using the following formula:

$$weight = \frac{\text{number of times term was mentioned}}{\text{total number of term mentions}}$$

[3] Available at: http://linkedinprofileservice.co/linkedin-profile-tips-advice/linkedin-skills-list/

For example, if there are a total of 4 term mentions and 'Linguistics' has been mentioned twice its weight will be 0.5.

Terms with only a single mention are discarded before weights are calculated. This decision was taken in order to minimise noise in the form of outlying items.

4.5 Comparison of representations

A term's weight in the LinkedIn or Twitter term lists generated by this analysis is directly related to the total number of term mentions in that list. As this total can differ between LinkedIn and Twitter, comparisons between term lists cannot be made using weights. Ranks are used instead.

For RQ1 the relative ranks of terms that appear in both the Twitter and LinkedIn term lists are compared.

For RQ2, only Twitter terms whose rank is equal to or higher than the lowest ranked term in the LinkedIn term list are recommended. For example, if the user's LinkedIn term list contains six ranks, only Twitter terms of rank six or higher are recommended. If no terms were found in the user's LinkedIn profile, only the first-ranked Twitter interest(s) is recommended.

5 Implementation

This section describes the implementation of the analysis described in the previous section.

5.1 Information Collected

The user's 1000 most recent tweets are collected using Twitter's public RESTful API[4]. The following information is collected using LinkedIn's public RESTful API[5]:

1. The user's summary
2. The user's skill, interest and course lists
3. The user's textual descriptions of their educational and professional experience.
4. The textual descriptions of LinkedIn groups to which the user belongs.

[4] https://dev.twitter.com/rest/public

[5] https://developer.linkedin.com/docs/rest-api

5.2 Term selection from resources

In WordNet, the hyperonymy relation links a noun synset to its parent. Analogously to the Swedish FrameNet++ lexical framework described by Forsberg and Borin (2014), in this study the Simple Knowledge Organization System (SKOS)[6] 'broader' relation is used as a DBpedia equivalent to hyperonymy.

5.2.1 WordNet

Extended WordNet Domain labels are used as terms. A keyword is linked to a WordNet synset and the domain label for this synset as well as the domain labels for its hyperonyms are obtained.

The '*factotum*' domain is not considered. This label is assigned to synsets to which no other label could be assigned, and thus has no specificity.

5.2.2 DBpedia

DBpedia category labels are used as terms. A keyword is linked to a DBpedia URI. This URI is then linked to the DBpedia category bearing the same label using the Dublin Core[7] 'subject' relation. If no such category exists, this means this URI content did not meet the criteria required to be given its own category[8]. In this case, all categories related to the URI by the 'subject' relation are obtained. Parent categories are identified through the SKOS 'broader' relation, and their labels are obtained.

The DBpedia category 'Main topic classifications' is not considered as it is a table of contents for other categories. Similarly, DBpedia categories such as 'Wikipedia categories named after information technology companies of the United States' are not considered as these refer specifically to the way in which the Wikipedia hierarchy is organised, rather than the concepts in it.

[6] http://www.w3.org/2004/02/skos/

[7] http://dublincore.org/

[8] Guidelines for creating Wikipedia categories: http://en.wikipedia.org/wiki/Wikipedia:Categorization

5.3 Texts

Texts (i.e. tweets, LinkedIn descriptions) are parsed, and the WordNet and DBpedia databases accessed, using the Pattern NLP library for Python (Smedt & Daelemans, 2012).

5.3.1 Tweet Preprocessing

Before tweets are analysed, they are preprocessed as follows:

- The word 'RT' appearing at the beginning of a tweet (indicating the tweet is a retweet) is removed.
- Characters repeated consecutively more than twice are replaced with two consecutive characters (e.g. 'gooood' becomes become 'good') as in (Vosecky et al., 2013).
- For hashtags the '#' symbol is removed and the tag is split using the capital-letter rule described in (Hauff & Houben, 2011). For example, '#MachineLearning' becomes 'Machine Learning'.

5.3.2 Corpora Used

In applying Magnini et al's approach, separate corpora are used for processing tweets and for processing LinkedIn textual data. For the former, a 36.4 million-word tweet corpus is used. For the latter the ~300 million-word blog corpus compiled by Schler and Koppel (2006) is used. This corpus is deemed suitable given both its size and the fact that it contains posts on a wide variety of topics, for example: 'Real Estate', 'Arts', 'Education', 'Engineering', 'Law' etc.

5.3.3 Modifications

This section describes modifications made to the analysis described in the 'Method' Section.

Magnini et al report that a context of at least 100 words should be used to disambiguate a word (50 words before the word and 50 words after). For tweets, as such a context is not available, the whole tweet is used.

The following modifications were made due to analysis time constraints.

Only links of 550 words or lower were analysed. This decision was taken with reference to the following quote from a Reuters blog post on the issue of ideal article length: 'Reuters editors see stories that exceed 500 or 600 words as indistinguishable from "Gravity's Rainbow")' (MacMillan, 2010).

In the DBpedia approach, creating feature vectors for the Naïve Bayes approach proved unworkable. Consider for example the term 'Xbox', which appears as a keyphrase in 4008 articles. To apply the Naïve Bayes approach, the following information would have to be gathered for each occurrence: 'the current word and its part-of-speech, a local context of three words to the left and right of the ambiguous word, the parts-of-speech of the surrounding words, and a global context implemented through sense specific keywords determined as a list of at most five words occurring at least three times in the contexts defining a certain word sense.'(Mihalcea and Csomai, 2007). This proved to be prohibitively expensive in terms of time taken. Thus, only Lesk's algorithm is used to disambiguate keywords in the DBpedia approach. However, the results reported by Mihalcea and Csomai for WSD using Lesk's algorithm alone are higher than approaches such as Agirre and Soroa (2009) and Gomes et al (2003). Thus, disambiguation quality is preserved.

6 Evaluation

Eight users participated in the evaluation. Participants were identified by two means: An email circulated in the research group in which the authors work; Tweeting at Twitter users from the university in which the authors work

The study is split into two sessions. In the first session, the user logs in to their Twitter and LinkedIn accounts. Their data is then collected and analysis performed. In the second session, the user is asked to subjectively evaluate DBpedia and Extended WordNet Domains with regard to each research question.

It must be noted that the number of participants in this experiment was quite small. As such, it would be unwise to make strong inferences from the results reported below.

6.1 RQ1 Procedure

Term comparisons are represented similarly to the study performed by Holland et al. (2003). Holland et al. represent user preferences for particular products in the format 'A is better than B'.

For RQ1, the user is shown a series of assertions about the relative ranks of terms that appear in both their Twitter and LinkedIn term lists. For example, if 'Linguistics' is the third ranked term in the user's LinkedIn term list but the fifth ranked term in their Twitter term list, the user is shown a statement asserting that Linguistics has less prominence in their tweets than in their LinkedIn profile. The user can answer affirmatively or negatively to each assertion. However, if the analysis has incorrectly identified a term, the user can indicate this instead of responding. They can also indicate that the term denotes an area that was of interest to them, but is not anymore.

If a term appears in the user's LinkedIn term list but not in their Twitter term list, the user is shown a statement asserting that the term has less prominence in their tweets than in their LinkedIn profile. If the user's LinkedIn term list is empty - as occurred with one user whose LinkedIn profile was sparse – no comparisons are made.

6.2 RQ2 Procedure

The approach for this question is similar to that adopted by Lee and Brusilovsky (2009). Lee and Brusilovsky use user judgments to evaluate the quality of the recommendations generated by their system. However, in Lee and Brusilovksy's study a Likert scale is used whereas in this study a multiple- choice format is used.

The user is shown a series of recommendations for their LinkedIn profile. The user can answer affirmatively or negatively to each recommendation. Alternatively, they can indicate that although the term denotes an area of interest to them they would not add it to their LinkedIn profile. This could be because they do not want to list the term on their professional profile or they do not feel sufficiently confident in their knowledge of the subject the term denotes. They can also indicate that the term denotes an area that was of interest to them, but is not anymore.

For the DBpedia approach, terms in the format 'Branches of X' are presented to the user as 'X', as these pages contain lists of sub-disciplines. For example, 'Branches of Psychology' becomes 'Psychology'. Similarly, terms in the format 'X by issue' are presented to the user as 'X'.

A user score is calculated for each lexical resource, with each research question contributing 50%. The scores from each user are then aggregated to give a final score for each resource.

7 Results

Table 1 illustrates the scores for each research question, while Table 2 shows error values. Extended WordNet Domains and DBpedia Categories are denoted using the acronyms EWND and DBC respectively. The figures in the tables are rounded.

Table Descriptions

Table 1

- ScrRQ1 – RQ1 score. The ratio of the number of correct comparisons to the total number of comparisons made.
- ScrRQ2 – RQ2 score. The ratio of the number of correct recommendations to the total number of recommendations made.

	ScrRQ1	ScrRQ2	Total	Total (all recs.)
EWND	40	30	35	54
DBC	62	51	53	70

Table 1. RQ1 and RQ2 Score percentages

	RQ1TermErr	ErrRQ2	RQ1Past	RQ2Past
EWND	29	30	4	1
DBC	16	14	0	0

Table 2. Error rate percentages

- Total – Obtained by adding the previous two scores together and dividing by 2.
- Total (all recs.) – The total including recommendations that were correct but which the user would not add to their LinkedIn profile.

Table 2

- Rq1TermErr – The percentage of prominence comparisons made containing incorrectly identified terms.
- ErrRQ2 – The percentage of incorrectly recommended terms.
- RQ1Past – The percentage of prominence comparisons made containing past interests.
- RQ2Past – The percentage of recommendations made containing past interests.

8 Discussion

The Extended WordNet Domains approach shows almost twice the percentage of incorrectly identified terms than the DBpedia approach. It also shows more than twice the percentage of incorrect recommendations. A reason for this can be found by examining the Extended WordNet Domains hierarchy. For example, consider the word 'law'. One of the possible synsets for this word defines it as 'the collection of rules imposed by authority'. The domain label for this synset is 'law'. The hyperonym for this synset is 'collection' whose definition is 'several things grouped together or considered as a whole'. The domain label for this synset is 'philately'. This directly contradicts the original WordNet Domains hierarchy, in which 'law' is a subclass of 'social science'.

The small number of study participants notwithstanding, the low error figures in the DBpedia approach look promising with regard to the task of profile aggregation. Abel et al. find that 'Profile aggregation provides multi-faceted profiles that reveal significantly more information about the users than individual service profiles can provide' (2010). Thus, a method that can accurately compare and combine information from a user's different profiles has value.

The marked difference between the 'Total' and 'Total (all recs.)' columns in Table 1 is also noteworthy. This indicates that there are certain subjects the study participants intended for Twitter, but not for LinkedIn.

One aspect of this study in need of improvement is the prominence comparisons (RQ1). During this part of the experiment, some participants said that they could not be sure about the relative weights of individual subject areas in their tweets and LinkedIn profile. However, in this case users were instructed to answer negatively so as not to artificially inflate scores. One way of overcoming this problem could be to generate ranked term lists for each profile and ask the user to subjectively evaluate each list separately.

9 Conclusion

This paper described a comparison between the Extended WordNet Domains and DBpedia lexical resources. The comparison took the form of an investigation of the ways in which users represent their interests and knowledge through their LinkedIn profile with the way they represent these characteristics through their tweets. In a user study with 8 participants the DBpedia category labels performed better than the WordNet Domain labels with regard to both research questions investigated.

10 Credits

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) in the ADAPT Centre (www.adaptcentre.ie) at Trinity College, Dublin.

11 References

Fabian Abel, Nicola Henze, Eelco Herder and Daniel Krause. (2010). Interweaving public user profiles on the web. *User Modeling, Adaptation, and Personalization*, pp. 16–27.

Fabian Abel, Eelco Herder, Geert-Jan Houben, Nicola Henze, and Daniel Krause (2012). Cross-system user modeling and personalization on the Social Web. *User Modeling and User-Adapted Interaction*, pp. 169–209.

Eneko Agirre, and Aitor Soroa. (2009). Personalizing pagerank for word sense disambiguation. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 33–41.

Lars Borin and Markus Forsberg. (2014). Swesaurus, or The Frankenstein Approach to Wordnet Construction. *Proceedings of the Seventh Global Wordnet Conference*, pp. 215–223.

Jeff Bullas. (2015). *5 insights into the Latest Social Media Facts, Figures and Statistics*. [Online] Available at: http://www.jeffbullas.com/2013/07/04/5-insights-into-the-latest-social-media-facts-figures-and-statistics/. [Accessed 25 March 2015]

Christian Chiarcos, John Mccrae, Petya Osenova, and Cristina Vertan. (2014). Linked Data in Linguistics 2014 . Introduction and Overview. *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, pp. vii–xv.

Qi Gao, Fabian Abel and Geert-Jan Houben. (2012). GeniUS: generic user modeling library for the social semantic web. *Proceedings of the Joint International Semantic Technology Conference*, pp. 160–175.

Jim Giles. (2005). Internet encyclopaedias go head to head. *Nature*, Vol. 438, pp. 900–901.

Paulo Gomes, Francisco C. Pereira, Paulo Paiva, Nuno Seco, Paulo Carreiro, José Luís Ferreira., and Carlos Bento. (2003). Noun sense disambiguation with WordNet for software design retrieval. 16th Conference of the Canadian Society for Computational Studies of Intelligence, pp. 537–543.

Aitor González, German Rigau, and Mauro Castillo. (2012). A graph-based method to improve WordNet Domains. *Proceedings of the Computational Linguistics and Intelligent Text Processing Conference*, pp.17–28.

Claudia Hauff, and Geert-Jan Houben. (2011). Deriving Knowledge Profiles from Twitter. *Proceedings of 6th European Conference on Technology Enhanced Learning: Towards Ubiquitous Learning*, pp. 139–152.

Danielle H. Lee, Peter Brusilovsky. (2009). Reinforcing Recommendation Using Negative Feedback. *User Modeling, Adaptation, and Personalization*, pp. 422–427.

Yunfei Ma, Yi Zeng, Xu Ren, and Ning Zhong. (2011). User interests modeling based on multi-source personal information fusion and semantic reasoning. *Proceedings of theActive Media Technology Conference*, pp. 195–205.

Robert MacMillan. (2010). *Michael Kinsley and the length of newspaper articles*.[Online] Available at: http://blogs.reuters.com/mediafile/2010/01/05/michael-kinsley-and-the-length-of-newspaper-articles/. [Accessed 13 March 2015].

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Massimiliano Gliozzo. (2002). The role of domain information in word sense disambiguation. *Natural Language Engineering*, vol. 8, pp. 359 – 373.

Rada Mihalcea, and Andras Csomai. (2007). Wikify!: linking documents to encyclopedic knowledge. *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. pp. 233-242.

Till Plumbaum, Songxuan Wu, Ernesto William De Luca and Sahin Albayrak (2011). User Modeling for the Social Semantic Web. *Workshop on Semantic Personalized Information Management*, pp. 78–89.

Katharina Reinecke and Abraham Bernstein:. (2009). Tell Me Where You ' ve Lived , and I ' ll Tell You What You Like : Adapting Interfaces to Cultural Preferences. *User Modeling, Adaptation, and Personalization*, pp. 185–196.

Ellen Riloff. (1999). Information Extraction as a Stepping Stone toward Story Understanding. *Understanding Language Understanding: Computational Models of Reading*, pp. 1–24.

Jonathan Schler, Moshe Koppel, Shlomo Argamon and James W. Pennebaker. (2006). Effects of Age and Gender on Blogging. *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*. pp. 199-205

Tom De Smedt and Walter Daelemans. (2012). Pattern for Python. *Journal of Machine Learning Research*, vol. 13, pp. 2063–2067.

Jan Vosecky, Di Jiang, Kenneth Wai-Ting Leung and Wilfred Ng. (2013). Dynamic Multi-Faceted Topic Discovery in Twitter. *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 879–884.