Expertise Mining

Georgeta Bordea and Paul Buitelaar

Unit for Natural Language Processing
Digital Enterprise Research Institute
National University of Ireland, Galway
{name.surname@deri.org}

Abstract. In this paper we present our approach for expertise mining, addressing several research problems such as expertise topic extraction, expert finding and expert profiling. We propose a hybrid solution inspired by different research fields such as expert finding, competency management, terminology extraction, keyword extraction and concept extraction. We introduce an expertise benchmarking dataset gathered by exploiting information about workshop committee members and we present Saffron, a system designed to give an overview of research areas and experts and to assist users in finding skilled individuals.

1 Introduction

The problem of searching for experts in a particular domain instead of documents, as in information retrieval, has received a lot of interest mainly for its applications in the organisational environment. Finding the appropriate person to consult is a cumbersome task, especially in organisations with a large number of members. This task is increasingly difficult when experts are needed outside of the organisation or in a different field. Typically expert search has been modelled in the same manner as the information retrieval process [1–3]. The users look for experts in a collection of documents by giving a query with topics of interest, then the query is matched against the document collection in order to find experts. The assumption these approaches make is that users are not interested in a detailed profile of the experts.

Expertise mining is another direction of research that addresses the same problem while focusing not only on identifying experts of a field but also on extracting relevant expertise topics for a specific domain and on building expertise profiles for people. The extraction of expertise topics starts at lexical level by identifying domain terminology and document keywords, then these terms are clustered to form concepts at the semantic level. Therefore expertise mining overlaps with a number of research areas such as expert finding, terminology extraction, keyword extraction and concept extraction.

The rest of the paper is organised as follows. We present some related work on competency management and expert finding in section 2, and we describe our approach for expertise mining in section 3. Section 4 presents our experiments and evaluation, and section 5 describes two applications for expertise mining. We conclude this paper in section 6 where we also mention future developments.

2 Background

Competence management is a research topic from the area of knowledge management that is concerned with the "identification of skills, knowledge, behaviours, and capabilities needed to meet current and future personnel selection needs, in alignment with the differentiations in strategies and organisational priorities" [4]. In [5] the relations between people and skills are extracted as a network of associations, both people and competencies being handled as entities. Buitelaar [6] presents an approach based on a combination of linguistic patterns, information retrieval statistical methods and machine learning is investigated.

Ontology based competency management. Paquette [7] introduces a competency ontology where expertise is considered a performance level of competency. According to his work, expertise can be evaluated by using several performance indicators such as frequency, scope, autonomy, complexity and context. The system introduced in [8] relies on ontologies for competency management, proposing a method to approximately match requirements and building inference services that update and derive skills from manually annotated documents. Ontologies can be also used for teams composition with matchmaking services that bring together the skills demand and supply [9]. The major obstacle that these methods face is that they can not be applied for domains where an ontology of skills is not already built. An approach to build an ontology of competencies has been proposed in [10], making use of an already built domain ontology and of CRAI model (Competency, Resource, Aspect, Individual) defined by Harzallah [11].

Expert finding. Extensive work has been done for the task of expert finding using information retrieval techniques. In these approaches users look for experts in a collection of documents by giving a query with topics of interest. Proposed solutions match the users query against the document collection in order to find experts. Documents available on an organisation's intranet are used as evidence of expertise and each expert is assigned an aggregated document of all associated documents [3]. Query expansion from the top ranked candidate profiles prove to be successful for this type of solution [1]. Expertise sources analysed are usually documents available inside an organisation. An aggregation model that assembles various kinds of information such as personal descriptions, related documents, similar people is described in [12] defining a multinomial probability distribution over the different sources of information. Several other sources are taken into consideration such as the global web (e.g. web search, news search, academic search) [2] or a community-maintained artefact of lasting-value (e.g. Wikipedia) [13]. Social networks are exploited with the assumption that expertise can be estimated with centrality measures [14]. A query-independent method that completely ignores the knowledge areas of expertise is designed for analysing the links acquired from posts and replies of specialised forums [15].

Building expert profiles. The importance of the expert profiling task in developing an expert finding system as a distinct task has been discussed in [16] but without addressing the problem of discovery and identification of possible knowledge areas. Two profiling methods are introduced, one that represents a person's skills as a score over documents that are relevant given a knowledge area and a second one that estimates the profiling scores using keyword similarity of candidates and knowledge areas. The authors focus mainly on measuring the competency of a person in an already given area. An expert finding solution that acquires expert profiles using document topics (i.e, index terms, topic keywords) appears in [17]. The problem of identity resolution is also addressed, as opposed to previous studies that focus only on string-based persons names.

3 Expertise mining

In this section we define the concepts of expertise profile and expertise topic and we propose a method to discover expertise topics in the task of building topical profiles. Then we present a method to estimate the level of expertise of a person. An expertise topic is the lexical realisation of a knowledge area. The expertise profile of an individual is a ranked list of expertise topics along with the evidence that supports these results (e.g., a list of documents). The first stage of building an expert profile is to identify the possible expertise topics. The next stage is to identify the experts for each expertise topic (i.e. the individuals that have the highest level of knowledge or know-how about a concept of the domain).

3.1 Expertise topics extraction

We rely on a list of manually identified skill types that are used in the text to introduce an expertise topic. Hearst [18] developed the well known idea of using context patterns to extract relations. Consider for instance the following extracts from scientific articles in the field of computer science:

*...**analysis** of social networks...*
*...**algorithm** for keyword extraction...*
*...**approach** for ontology population...*
*...**method** for geographical information retrieval...*

In all four examples the expertise topic (e.g., "social networks", "keyword extraction", "ontology population", "geographical information retrieval") is introduced by a skill type (e.g., "analysis", "algorithm", "approach", "method"). Some of these skill types are valid for any scientific area (e.g. "approach", "method", "analysis", "solution") while other skill types are domain specific, e.g., for computer science "implementation", "algorithm", "development", "framework", for physics "proof", "principle","explanation" and for chemistry "law", "composition", "mechanism", "reaction", "structure". By adapting the list of skill types this method can be applied not only to scientific publications, but also to corpora gathered from the intranet of large organisations.

The syntactic description for a term [19] is analysed to discover candidate expertise topics in the context of a type of skill. Expertise topic patterns are defined by a sequence of part-of-speech tags, mainly a noun phrase but also proper nouns, commonly used as topic names in computer science (e.g. "Semantic Web", "Social Networks", "Rule-Based WSML"), cardinal numbers (e.g. "P2P systems") and gerunds (e.g. "ontology mapping", "data mining"). A combination of statistical measures is used to rank the candidate expertise topics, taking into consideration the relevance of an expertise topic.

Ranking. Longer expertise topics in terms of number of words are ranked higher, because they refer to a more specific concept and they tend to be mentioned less often. Keyphrases identified frequently as a candidate expertise topic are also ranked higher. We make a distinction between the frequency of an expertise topic occurring in the context of a skill type and the overall occurrence of an expertise topic. Therefore we define the score for an expertise topic as:

$$R_i = Tn_i * Fn_i$$

Where R_i is the score for the candidate expertise topic i, Tn_i is the normalized number of tokens (number of tokens divided by the maximum number of tokens of a phrase) and Fn_i is the normalized frequency of the term in the context of a skill type (frequency divided by the maximum frequency of a phrase).

Filtering. We used an external web search engine to filter out the expertise topics that are too specific or too general from the final result list. We considered that if a phrase has less than 5 hits on the web it is too specific for the document collection to be taken into consideration (e.g. “deciphering user intention”, “computing keyword prefixes”, “extracting meaningful and representative clusters”), or there was an extraction error during extraction (e.g. “diversi cation systems”, “flnding news articles”). Some of the errors appear while parsing text from PDF documents, while others appear because of incorrectly tagged part-of-speech elements. If an expertise topic has more than 10^9 hits on the web it is too general to be included in the final result set of expertise topics (e.g. “Internet Explorer”, “service”, “community”, “website”).

3.2 Expert profile construction

The expertise topics extracted from a publication are added to the expertise profiles for each of the researchers mentioned in the author list, considering that the authors of a publication are subject matter experts. No distinction is made based on the sequence of authors of a paper, assuming that all authors have the same level of expertise on the topics mentioned in the publication.

Each expertise topic mentioned in the researcher’s publications is assigned a measure of relevance computed using an adaptation of the standard information retrieval measure TF-IDF. The set of documents of a researcher is aggregated in a virtual document and the relevancy of an expertise topic is computed over this virtual document. After expertise topic extraction, expertise topic ranking, expertise topic filtering and researcher relevance scoring, each expertise topic in the expertise profile of a researcher is scored according to the formula:

$$S_{t,r} = R_t * tfirf_{t,r}$$

Where $S_{t,r}$ represents the score for an expertise topic t and a researcher r, R_t represents the rank computed in section 3 for the expertise topic t and $tfirf_{t,r}$ stands for the TF-IDF measure for the expertise topic t and the virtual document for researcher r that is computed by taking the overall occurrence of an expertise topic into consideration. Other factors can be used to measure the relevance of expertise topics such as links between researchers extracted from the co-citation graph or the co-authorship graph. The experience of a person for an expertise topic can be estimated by measuring the number of publications mentioning the expertise topic or by analysing their timeline.

More sophisticated methods for expert profiling can be envisioned by defining expertise as a performance level of competency. For instance an expert should have a good coverage of different expertise topics related to an area and should be able to apply his knowledge in other contexts.

4 Evaluation

4.1 Datasets

In order to evaluate our approach we use various data sets of scientific publications from several areas of computer science. The first dataset is a corpus of scientific publications from Semantic Web conferences[1] produced by the semantic web community and it consists of 846 papers and 2008 researchers from 11 semantic web conferences starting from 2006 to 2010. The ACL Anthology Reference Corpus[2] is a second much larger dataset produced by the computational linguistics community that contains 10921 scientific articles, published between 1965 to 2006 and that contains information about 9983 researchers. We also analysed a collection of articles published by researchers working in a web science research institute, DERI, NUI Galway, that contains 405 scientific publications and 362 researchers.

4.2 Benchmarking

We evaluate our results both at the expertise topic extraction level, by comparison with keyword extraction baselines and at the expert profile construction level by introducing a benchmark dataset.

Keyword extraction. The SemEval 2010 competition included a task targeting the Automatic Keyphrase Extraction from Scientific Articles [20]. Given a set of scientific articles participants are required to assign to each document keyphrases extracted from text. We participated in this task with an unsupervised approach for keyphrase extraction [21] that does not only consider a general description of a term to select candidates but also takes into consideration context information. The SemEval task organizers provided two sets of scientific articles, a set of 144 documents for training and a set of 100 documents for testing. Three sets of answers were provided: author-assigned keyphrases, reader-assigned keyphrases and combined keyphrases (combination of the first two sets). The participants were asked to assign a number of exactly 15 keyphrases per document.

All reader-assigned keyphrases are extracted from the papers, whereas some of the author-assigned keyphrases do not occur explicitly in the text. The traditional evaluation metric is followed, matching the extracted keyphrases with the keyphrases in the answer sets and calculating precision, recall and F-score. In both tables the column labels start with a number which stands for the top 5, 10 or 15 candidates. The characters P, R, F mean micro-averaged precision, recall

[1] Semantic Web Corpus: http://data.semanticweb.org

[2] ACL Anthology Reference Corpus: http://acl-arc.comp.nus.edu.sg

and F-scores respectively. For baselines, 1, 2, 3 grams were used as candidates and TF-IDF as features. The same conventions stand for table 2. In table 1 the keyphrases extracted by our system are compared with keyphrases extracted by an unsupervised method that ranks the candidates based on TF-IDF scores and two supervised methods using Naive Bayes (NB) and maximum entropy(ME) in WEKA[3]. Our performance (DERIUNLP) is well above the baseline in all cases.

Method	**5P**	**5R**	**5F**	**10P**	**10R**	**10F**	**15P**	**15R**	**15F**
TF-IDF	22%	7.5%	11.19%	17.7%	12.07%	14.35%	14.93%	15.28%	15.1%
NB	21.4%	7.3%	10.89%	17.3%	11.8%	14.03%	14.53%	14.87%	14.7%
ME	21.4%	7.3%	10.89%	17.3%	11.8%	14.03%	14.53%	14.87%	14.7%
DERIUNLP	27.4%	9.35%	13.94%	23%	15.69%	18.65%	22%	22.51%	22.25%

Table 1. Baseline and DERIUNLP Performance aver Combined Keywords

Expertise benchmark dataset. Further benchmarking on the more challenging expertise level is done by gathering data about workshop committee members, assuming that their selection is based on human judgement. Another advantage of using workshop data is that typically they are concerned with a focused subject and the main expertise topics can be manually extracted from the title of the workshop. The information about committee members was manually extracted from workshop description files. In this way we collected a benchmark dataset with workshops mentioned in the Semantic Web Corpus and a second benchmark dataset from the ACL corpus.

Method	**5P**	**5R**	**5F**	**10P**	**10R**	**10F**	**15P**	**15R**	**15F**
TFIRF	3.43%	2.32%	2.77%	3.23%	4.37%	3.71%	2.49%	5.05%	3.34%
YearCount	4.24%	2.87%	3.42%	4.55%	6.15%	5.23%	3.84%	7.79%	5.14%
Sindice	11.31%	7.65%	9.13%	9.39%	12.70%	10.80%	7.47%	15.16%	10.01%

Table 2. Performance of tfirf, number of years and number of web hits

The first dataset contains information about 102 workshops and we manually extracted 126 expertise topics from the workshop names. The second dataset is extracted for 248 workshops and contains 286 expertise topics. Table 2 presents the evaluation of three different factors for expert finding in terms of precision, recall and f-score. The *tfirf* factor is described in section 3, the *YearCount* factor represents the number of years an expertise topic was mentioned and the *Sindice* factor refers to information extracted using a Semantic Web search engine.

5 Applications

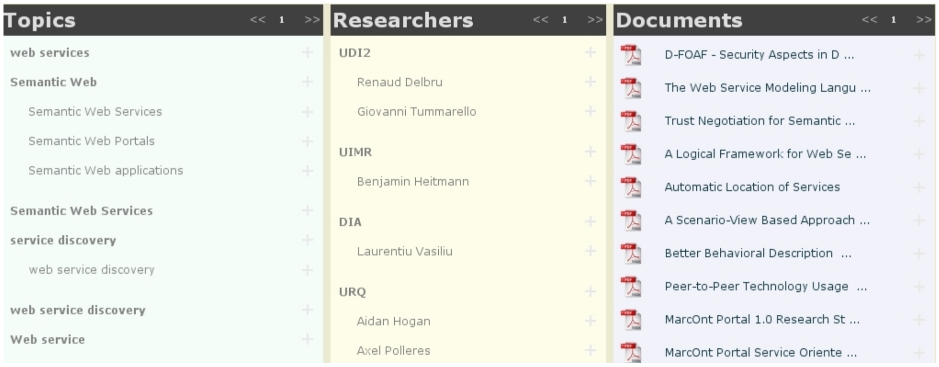
This section is concerned with describing two applications that make use of the information extracted with our approach.

[3] WEKA:http://www.cs.waikato.ac.nz/ml/weka/

5.1 Saffron. An expert profiling and expert finding system

Research organisations typically have a large number of research groups, projects, papers and members making it difficult to get insights into everything going on. It is not uncommon that researchers, newcomers as well as more established ones, are not fully aware of the research topics their colleagues are working on, with negative impact on research quality. Saffron addresses this problems by providing researchers with a deeper understanding of active or previous topics researched in their own organisation.

User interface. Figure 1 presents the user interface. The main part of the user interface consists of three columns, corresponding to three resource types: topics, researchers and documents. Each column displays the corresponding resources in a hierarchical view with bold parent items and with indented child items.



- **web services**
- **Semantic Web**
  - Semantic Web Services
  - Semantic Web Portals
  - Semantic Web applications
- **Semantic Web Services**
- **service discovery**
  - web service discovery
- **web service discovery**
- **Web service**

- **UDI2**
  - Renaud Delbru
  - Giovanni Tummarello
- **UIMR**
  - Benjamin Heitmann
- **DIA**
  - Laurentiu Vasiliu
- **URQ**
  - Aidan Hogan
  - Axel Polleres

- D-FOAF - Security Aspects in D ...
- The Web Service Modeling Langu ...
- Trust Negotiation for Semantic ...
- A Logical Framework for Web Se ...
- Automatic Location of Services
- A Scenario-View Based Approach ...
- Better Behavioral Description ...
- Peer-to-Peer Technology Usage ...
- MarcOnt Portal 1.0 Research St ...
- MarcOnt Portal Service Oriente ...

Fig. 1. The user interface of the Saffron system.

The initial view presents top ranked topics, researchers and documents. Topic ranks are passed from the extraction phase, while ranks for researchers and documents are calculated as sums of ranks of associated topics. Effectively, a user sees the most relevant topics within an institute, who are the experts on these topics and which are the top documents published about these topics. A user can select any of the visible topics, researchers or documents and the content of the three columns will change. Users who search for a particular topic or researcher are not constrained to visually search for it in the interface, instead a text based search facility is available. When a resource is found it will be selected as if the user clicked on it and the content of the other columns will be modified to display related resources.

System components. The expertise mining component is based on the GATE natural language processing framework [22]. Figure 2 presents the most important sub-components of the expertise mining module. We use the ANNIE IE system included in the standard GATE distribution for text tokenisation, sentence splitting and part-of-speech tagging. We annotate skill types using a gazetteer and expertise topics using regular expressions.

The expertise topics extracted with the GATE pipeline are stored into a relational database along with frequency information and metadata about the

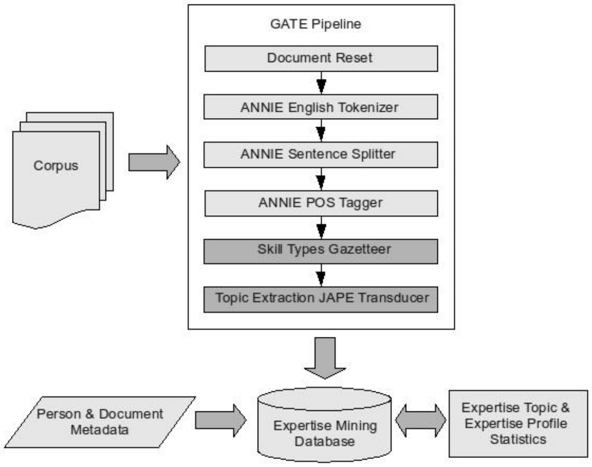

GATE Pipeline
Document Reset
ANNIE English Tokenizer
ANNIE Sentence Splitter
ANNIE POS Tagger
Skill Types Gazetteer
Topic Extraction JAPE Transducer
Corpus
Person & Document Metadata
Expertise Mining Database
Expertise Topic & Expertise Profile Statistics

Fig. 2. Architecture of the Expertise Mining module.

documents such as title and list of authors. The statistics component takes care of ranking and filtering the candidate expertise topics and of generating the expertise profiles for individual researchers accordingly. At any given point, a user can select any of the visible topics, subtopics, units, researchers, or documents. In such a case, the content of the three columns will change.

5.2 Conference and journal submission recommender

Another possible application for the knowledge extracted for expertise mining is a submission recommender system that builds a profile for a number of different conferences or journals. Scientific articles published in each conference and journal are gathered along with the topics from call for papers and information about the impact factor. The system extracts the expertise topics from an article and then suggests a list of appropriate events and periodicals. In this way researchers are assisted in their effort of keeping up to date with the best events in their field.

6 Conclusions

Although local context, such as cue phrases, is typically used in information extraction for identifying particular types of information, not much thought has been given to the semantic interpretation of such contexts. In this paper we presented our work on establishing this for the extraction task of Expertise Mining, a specialized task of text mining. In this case, expertise topics are typically introduced by a specific set of “skill types” (e.g., “methods”, “approach”, “analysis”), which we argue can be semantically organized according to the kind of expertise they introduce.

The method proposed here uses term extraction techniques (the syntactic description of a term), keyword extraction (TF-IDF) and contextual evidence (skill types) for the type of information we extract (expertise topics) to improve baseline statistical measures in expertise mining. Future work will include an algorithm for automatic extraction of the skills types for a domain and an analysis of the performance of each type of skill in the task of expertise mining.

We also plan to "semantify" the expertise topics and skill types by associating them with background knowledge available from the Linked Data[4] cloud. To achieve this we have to disambiguate those that refer to several Linked Data URIs (i.e., concepts) using word sense disambiguation techniques. Another problem is that different expertise topics can refer to the same concept, therefore we will explore their similarity using clustering techniques. Finally, we also plan to investigate potential relations between expertise topics based on cooccurrence analysis. We will investigate how the proposed methods, especially the skill types can be adapted to different scientific domains by analysing the collection of publications produced in different departments of a university.

7 Acknowledgements

This work is supported in part by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

References

1. Macdonald, C., Ounis, I.: Voting for candidates: adapting data fusion techniques for an expert search task. In: CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management, New York, NY, USA, ACM (2006) 387–396
2. Serdyukov, P., Rode, H., Hiemstra, D.: Exploiting sequential dependencies for expert finding. In: SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2008) 795–796
3. Craswell, N., Hawking, D., Vercoustre, A.M., Wilkins, P.: P@noptic expert: Searching for experts not just for documents. In: In Ausweb. (2001) 21–25
4. Draganidis, F., Metzas, G.: Competency based management: A review of systems and approaches. Information Management and Computer Security **14**(1) (2006) 51–64
5. Zhu, J., Goncalves, A.L., Uren, V.S., Motta, E., Pacheco, R.: Mining web data for competency management. In: In Proc. of Web Intelligence (WI 2005, IEEE Computer Society (2005) 94–100
6. Buitelaar, P., Eigner, T.: Topic extraction from scientific literature for competency management. In: Personal Identification and Collaborations: Knowledge Mediation and Extraction (PICKME2008). (2008)
7. Paquette, G.: An ontology and a software framework for competency modeling and management. Educational Technology & Society **10**(3) (2007) 1–21
8. Sure, Y., Maedche, A., Staab, S.: Leveraging corporate skill knowledge – from proper to ontoproper. In: Proceedings of the third international conference on practical aspects of knowledge management. (2000) 30–31
9. Colucci, S., Noia, T.D., Sciascio, E.D., Donini, F.M., Piscitelli, G., Coppi, S.: Knowledge based approach to semantic composition of teams in an organization. In: SAC '05: Proceedings of the 2005 ACM symposium on Applied computing, New York, NY, USA, ACM (2005) 1314–1319

[4] Linked Data: http://linkeddata.org

10. Posea, V., Harzallah, M.: Building a competence ontology. In: proc. of the workshop Enterprise modelling and Ontology of the International Conference on Practical Aspects of Knowledge Management (PAKM 2004). (2004)
11. Harzallah, M., Berio, G.: Competence modeling and management: A case study. In: ICEIS (3). (2004) 350–358
12. Zhang, W., Chang, L., Ma, J., Zhong, Y.: Aggregation models for people finding in enterprise corpora. In Karagiannis, D., Jin, Z., eds.: KSEM. Volume 5914 of Lecture Notes in Computer Science., Springer (2009) 180–191
13. Demartini, G., Firan, C.S., Iofciu, T., Krestel, R., Nejdl, W.: A model for ranking entities and its application to wikipedia. In Baeza-Yates, R.A., Jr., W.M., Santos, L.A.O., eds.: LA-WEB, IEEE Computer Society (2008) 29–38
14. Campbell, C.S., Maglio, P.P., Cozzi, A., Dom, B.: Expertise identification using email communications. In: CIKM, ACM (2003) 528–531
15. Zhang, J., Ackerman, M.S., Adamic, L.A., Nam, K.K.: Qume: a mechanism to support expertise finding in online help-seeking communities. In Shen, C., Jacob, R.J.K., Balakrishnan, R., eds.: UIST, ACM (2007) 111–114
16. Balog, K., de Rijke, M.: Determining expert profiles (with an application to expert finding). In: proc. of the International Joint Conferences on Artificial Intelligence (IJCAI 2007). (2007)
17. Jung, H., Lee, M., Kang, I.S., Lee, S., Sung, W.K.: Finding topic-centric identified experts based on full text analysis. In Zhdanova, A.V., Nixon, L.J.B., Mochol, M., Breslin, J.G., eds.: FEWS. Volume 290 of CEUR Workshop Proceedings., CEUR-WS.org (2007) 56–63
18. Hearst, M.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: Proceedings of the 14th International Conference on Computational Linguistics. (1992) 539–545
19. Mima, H., Ananiadou, S., Matsushima, K.: Terminology-based knowledge mining for new knowledge discovery. ACM Trans. Asian Lang. Inf. Process. **5**(1) (2006) 74–88
20. Kim, S.N., Medelyan, A., Kan, M.Y., Baldwin, T.: SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. In: Proceedings of the ACL 2010 Workshop on Evaluation Exercises on Semantic Evaluation (SemEval 2010). (2010)
21. Bordea, G., Buitelaar, P.: DERIUNLP: A Context Based Approach to Automatic Keyphrase Extraction. In: Proceedings of the ACL 2010 Workshop on Evaluation Exercises on Semantic Evaluation (SemEval 2010). (2010)
22. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. (2002)