Identify Experts from a Domain of Interest

Adrian Iftene, Bogdan Luca, Georgiana Cărăușu, Madălina Merchez

UAIC: Faculty of Computer Science, "Alexandru Ioan Cuza" University,
General Berthelot, 16, 700483, Iasi, Romania
{adiftene, bogdan.luca, georgiana.carausu, madalina.merchez}@infoiasi.ro

Abstract. User networks are beginning to be increasingly difficult to manage because of the large volume of information which is circulated within them. For example, in the Yahoo!Answers network, the large number of questions makes the identification of an expert, who would be the most suited to answer a question, a long-lasting process (currently this process is semi-automatic). This paper proposes an automatic identification method of a human expert, who would be the most suited to answer a question from a certain user of Yahoo network.

Keywords: Yahoo!Answers, WordNet, Google Translate

1 Introduction

This paper deals with the problem of identifying a domain expert in a multilingual context of search offered by social networks. The problem is topical and solving it is of a great interest in the online communities. Therefore, among the exercises of the CLEF 2010[1] assessment there was an exercise especially for this purpose CriES[2]. This exercise's main purpose was to identify experts in the context of multilingual search. This challenge is related to the problem of human expert search, i.e. those members of online communities, which can solve new problems, can answer questions, or requests for support from social multilingual networks.

For the evaluation exercise, the organizers provided a subset of a collection from Yahoo!Answers[3] containing 60 questions in 4 different languages: English, French, German and Spanish for which we had to find experts. The original file of over 12 GB of data was processed with a processing tool provided by the organizers. Following this processing we obtained a file with only 204 domains of interest of approximately 800 MB and a file containing a digraph of questions.

The nodes of the digraph represent the IDs of the users who asked questions, the IDs of the users who responded, and the edges represent the question's domain.

The last file we obtained was a file with 60 questions for which we had to identify the expert that would help us in getting a response.

[1] CLEF 2010: http://clef2010.org/
[2] CriES: http://www.multipla-project.org/cries:start?redirect=1
[3] Yahoo!Answers: http://answers.yahoo.com/

2 Existing Work

In [5], Sorg and Cimiano represent the documents as vectors in the Wikipedia articles space, using Tf-idf measure[4] to determine how "important" a Wikipedia article for a specific word is.

Later, in 2009 the same authors in [6] present a classification method based on multilingual links. Their approach works for language pairs for which there exists a substantial number of multilingual links.

In [4] the authors present how they used an approach based on explicit semantic analysis in processing steps automatic language identification and how they used different strategies to achieve the final rankings.

[1] presents how search models can be compared based on explicit concepts with models based on latent concepts using in training process parallel multilingual collections JRC-Acquis[5] and Multext[6].

3 System Components

Our system is composed of several modules dealing with various types of processing. The most important components of the system deal with eliminating unimportant words, obtaining synonyms for English words and with translation in and from English of initial words of the user question. Next we present the main components of this system presented in Figure 1.

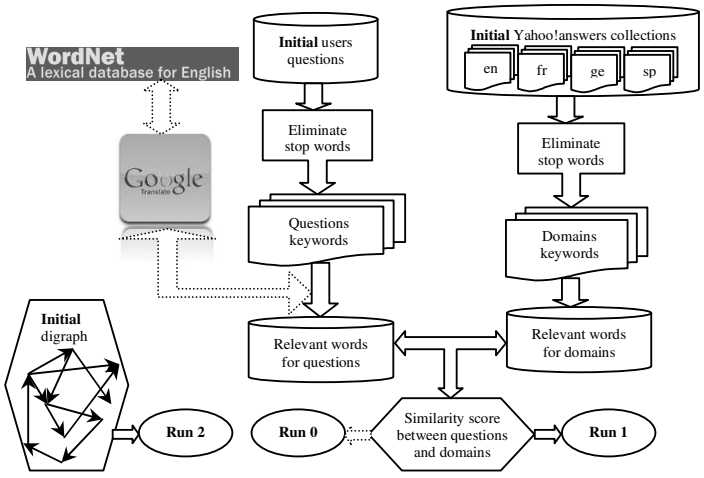

Fig. 1. UAIC system main components

[4] Tf-idf measure: http://en.wikipedia.org/wiki/ Tf%E2%80%93idf
[5] JRC-Acquis: http://wt.jrc.it/lt/Acquis/
[6] Multext: http://nl.ijs.si/ME/

Getting Keywords and Eliminating Irrelevant Words

For each domain of interest for which we must obtain a list of experts, we divided the information from the tags <title> and <description> in a list of words. From that list we removed the irrelevant words for the language which includes that domain, such as "the", "and", "is" for English, "je", "la", "le" for French, etc. Thus for each domain we added another tag <keywords> containing the list of relevant words for the domain.

Obtaining the Synonyms Lists

Given the list of keywords for each topic, we used Google Translate[7] and we translated the keywords into English. Then using the English version of WordNet[8] we obtained the list of synonyms for the translated words. After this step we used Google Translate again and we translated the synonyms in the original language. Thus for each domain we added another tag, <synonyms> where we put the synonyms of the keywords obtained from the previous step.

Grouping the Questions and Answers in Domains

To speed processing on each domain, we decided to divide the original XML which contained all domains with the questions and answers (approximately 800 MB) in smaller XMLs, which are easier to process. Thus, for each tag containing the question and the answers, we determined which category it belongs to and we put it in a new XML named after the category's name. Finally the original file was divided into 204 smaller files.

4 Submitted Runs

Using various combinations of modules and components we built 3 runs that we sent to the organizers of this exercise. See Figure 1 for more details.

Run 0

Initially, in our opinion, this was supposed to be the best result. In this case we consider word synonyms in the search process. Our assumption was that this type of search will get better results, as it has already been shown in previous works [2], [3] and [8]. This run was obtained through the following steps:

[7] Google Translate: http://translate. google.com/
[8] English WordNet: http://wordnet.princeton.edu/

- **Step 1**: for each question for which we had to find it's expert we determined which category it belongs to (for that we used the <category> tag) and we used in the following stages of processing the corresponding file obtained in the pre-processing stage.
- **Step 2**: in the second step we calculated a similarity score between the question and the question-answer elements existing in domain files. For that we consider the added tags, <keywords> and <synonyms>.
    - **Step 2.1**: the similarity score between the current question and a question-answer pair from a domain file increased by two points for each word from the question that belongs to the <keywords> tag from the topic or by one point for each word from the question that belongs to the <synonyms> tag from the topic.
    - **Step 2.2**: in the second stage we summed the obtained scores for each person who answered lots of questions.
- **Step 3**: in the end we considered as experts only the first 10 users in descending order of the amount scores obtained in the previous step.

Run 1

The second run follows the same steps presented above, the only difference being related to the calculation of score in Step 2.1. In this run we didn't take account of the changes in scores due to <synonyms> tags.

Run 2

For our third run we used the digraph provided by Yahoo, in which the nodes were user IDs and the edges signified that a user answered to another user's question, the question belonging to a certain domain. In this case, we considered for each user the number of answers given by him in a given domain as the number of the edges with questions in that domain to which that user answered. For that we consider only the <category> tag from the file with questions and the number of answers given by users in a given domain. Finally the expert ranking was obtained by the descending order of the user scores.

5 Results

Our official results are presented in Table 1 and they are taken from [7] (where *P@10* represents "*precision at cut-off level 10*" and *MRR* represents "*Mean Reciprocal Rank*").

Table 1: Results of UAIC's runs

Run Id	Characteristics	Strict		Lenient	
		P@10	**MRR**	**P@10**	**MRR**
0	We eliminate stop words and we consider relevant keywords and their synonyms (using Google Translate and English WordNet)	0.52	0.80	0.82	0.94
1	We eliminate stop words and we consider only relevant keywords	0.47	0.77	0.77	0.93
2	We consider only the digraph provided by Yahoo	**0.62**	**0.84**	**0.83**	**0.94**

Contrary to our expectations the best result was obtained for Run 2, where we consider only the digraph in order to identify the experts. Obviously, the score of Run 0, where we consider keywords and their synonyms in the process of calculation the score is better than the score of Run 1, where we consider only keywords. In the future, we must conduct a more detailed investigation of the evaluation results in order to better understand what happened with the results of Run 2.

6 Conclusions

In this paper we presented our group's participation in the CriES 2010 exercise from CLEF 2010. Based on Google's translation service and using the English WordNet word synonyms we got three runs that we sent to the organizers of this evaluation exercise. Run 2 and Run 0 were our best runs and they had a very good classification (see [7] for more details).

In the future we also want to use the multilingual features of the collection offered by the competition's organizers, because we believe that this area can bring significantly improved results to our system.

Acknowledgements. The research presented in this paper was funded by the Sector Operational Program for Human Resources Development through the project "Development of the innovation capacity and increasing of the research impact through post-doctoral programs" POSDRU/89/1.5/S/49944. The authors of this paper thank the colleagues from the B6 group, II year, Faculty of Computer Science Iasi, for the help offered in this project.

References

1. Cimiano, P., Delft, T.U., Schultz, A., Sizov, S., Sorg, P., Staab, S. Explicit vs. Latent Concept Models for Cross-Language Information Retrieval. IJCAI'09: Proceedings of the

21st international joint conference on Artificial intelligence, San Francisco, CA, USA, (2009)
2. Mihalcea, R., Moldovan, D. Semantic Indexing using WordNet Senses. In Proceedings of ACL Workshop on IR & NLP, Hong Kong, October (2000)
3. Rosso, P., Molina, A., Pla, F., Jiménez, D., Vidal, V. Information Retrieval and Text Categorization with Semantic Indexing. CICLing 2004, Pp. 596-600 (2004)
4. Sorg, P., Braun, M., Nicolay, D., Cimiano, P. Cross-lingual Information Retrieval based on Multiple Indexes. Working Notes for the CLEF2009, 30 September - 2 October, Corfu, Greece (2009)
5. Sorg, P., Cimiano, P. Enriching the Crosslingual Link Structure of Wikipedia - A Classification-Based Approach. AAAI2008, Institute AIFB, University of Karlsruhe, D-76128 Karlsruhe, Germany (2008)
6. Sorg, P., Cimiano, P. An Experimental Comparison of Explicit Semantic Analysis Implementations for Cross-Language Retrieval. Working Notes for the CLEF2009, 30 September - 2 October, Corfu, Greece (2009)
7. Sorg, P., Ciminao, P., Sizov, S. Overview of the Cross-lingual Expert Search (CriES) Pilot Challenge. Working Notes of the CLEF 2010 Lab Sessions, 20-23 September, Padua, Italy (2010)
8. Voorhees, E. M. Using WordNet to disambiguate word senses for text retrieval. SIGIR'93. Pp. 171-180, ACM, New York, USA (1993)