Discovering and Merging Keyword Senses using Ontology Matching*

Mauricio Espinoza**, Raquel Trillo, Jorge Gracia, and Eduardo Mena

IIS Department, Univ. of Zaragoza, María de Luna 1, 50018 Zaragoza, Spain
{mespinoz,raqueltl,jogracia,emena}@unizar.es
WWW home page: http://sid.cps.unizar.es

Abstract. During the last years we are witnessing how the use of keywords has become the standard input when searching the Web. As opposite to the syntactic searches performed by traditional web search engines, the current research challenge is a semantics-guided information retrieval. The increasing pools of ontologies available on the Web can help to discover the semantics of user keywords and this information is priceless for many tasks, including new semantic search engines.
In this paper we propose a system that takes as input a list of keywords provided by the user and discovers their possible meanings by consulting the knowledge represented by many (heterogeneous and distributed) ontologies. These keyword senses are semantically enriched with the synonym terms found during the ontology matching process: A synonymy measure based on statistics techniques and ontological similarity is used to integrate senses that are similar enough.

Keywords: Ontology matching for information integration

1 Introduction

Although keyword-based search is a widely used technique for information retrieval, traditional techniques do not consider the specific semantics assigned by the user: the same keywords can be used by different users with the purpose of accessing to different information. Furthermore, the syntactic-based search engines are very influenced by the enormous amount of information about popular issues on the Web, i.e., the keyword "java": Java as programming language eclipses the rest of possible senses (the Indonesian island, a coffee plant, different US cities, etc). However, ontologies (which offer a formal, explicit specification of a shared conceptualization [5]) can be used to make the semantics of user keywords explicit without ambiguity. The more ontologies consulted, the more chances to find the semantics assigned to keywords by the user.

In this paper, we propose a system that takes as input a list of plain keywords provided by the user, discovers their semantics in run-time and obtains a

* This work is supported by the CICYT project TIN2004-07999-C02-02.
** Work supported by a grant of Santander Central Hispano & University of Zaragoza.

list of senses extracted from different ontology pools; it deals with the possible overlapping among senses. The main steps of our approach are summarized in the following:

1. *Extraction of Keyword Senses.* First, the user keywords are normalized by a preprocessing step (e.g., rewriting them in lowercase, removing hyphens, etc.), and in order to discover the semantics of the user keywords, the system accesses to the shared knowledge stored in different ontology pools available on the Web. The extracted senses are semantically enriched with the ontological senses of their synonyms (which are obtained from the ontology pool), whenever the system evaluates that the synonym senses matches to the semantics of the corresponding keyword sense.
2. *Alignment of Senses.* This process uses an incremental algorithm for the alignment of the different keyword senses in order to remove the possible semantics redundancy among them. Senses are merged when the estimated synonymy probability between them is above a certain threshold. The synonymy measure combines a standard string distance metric with a structural similarity measure that is based on vector space techniques. Thus the result is a set of *different* possible senses for each user keyword.

For efficiency purposes, the system uses sampling and other statistic techniques, as well as parallel processing, whenever possible. The output of our system can be the input for a disambiguation process across keywords [4] or used to retrieve data once the keyword semantics is known.

The rest of this paper is as follows. In Section 2 we show how the possible senses of each keyword are obtained and semantically enriched with their synonym senses. In Section 3 we describe the algorithm that computes the synonymy probability in order to integrate senses when a certain threshold is achieved. Finally, conclusions and future work appear in Section 4.

2 Extraction of Keyword Senses

In this section we provide the details that show the contribution of this paper in the task of automatically retrieving the possible senses for a set of user keywords. In order to find the ontological terms that match those keywords, the system accesses to Swoogle [2], other remote lexical resources as WordNet [8] and other ontologies not indexed by Swoogle are used as well. We advocate using a pool of ontologies instead of just a single one, like WordNet (as many works do [6, 7]), because many technical or subject-specific senses cannot be found in WordNet.

The system builds a sense for each URI obtained with the information retrieved from matching terms in the ontology pool [1]. In our approach, a sense of a keyword k, denoted by s_k, is a tuple $s_k = <s,\ grph,\ descr,\ pop,\ syndgr>$, where s is the list of synonym names[1] of keyword k, *grph* describes the sense

[1] To extract from an ontology the synonyms of a class, property or individual, the primitives *equivalentClass*, *equivalentProperty* and *sameIndividualAs* are used, respectively.

s_k by means of the hierarchical graph of hypernyms and hyponyms of synonym terms found in one or more ontologies, *descr* is a description in natural language of such a sense, and *pop* and *syndgr* measure the degree of popularity of this sense (*pop* is the number of times it appears in the ontology pool and *syndgr* is the integrated percentage of synonymy degree). Thus, senses are built with the information retrieved from matching terms in the ontology pool [1].

As matching terms could be ontology classes, properties or individuals, three lists of possible senses are associated with each keyword k: S_k^{class}, S_k^{prop} and S_k^{indv}. In Figure 1 we show an example of some senses found in the ontology pool for the user keyword "star". The system finds in WordNet two matchings of keyword "star" as concept/class ($s1$ and $s2$), and one matching in the Travel Ontology[2] as property of class "hotel" ($s3$). Notice that each sense is initialized with a popularity=1 and a synonymy degree=1.

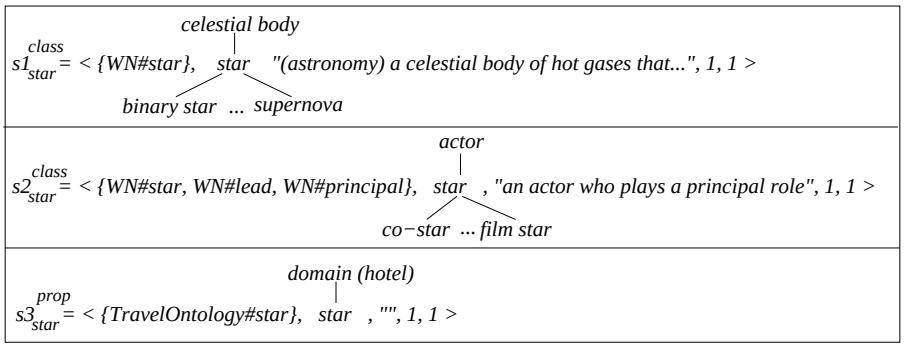

Fig. 1. Some senses of keyword "star" extracted from the ontology pool

Each keyword sense is enhanced incrementally with the synonyms terms extracted from the ontology pool. Therefore our system takes advantage of the shared ontologies available on the Web and semantically enriches the keyword senses with senses extracted from their synonyms. The synonym names are stored in the sense structure shown before, which gets upgraded everytime the sense is integrated with a (very similar) sense coming from other ontology. In order to evaluate the semantic similarity between the sense of a keyword and their synonyms, the system performs a *sense alignment* process (detailed in Section 3) which determines whether the semantics of the keyword sense and each synonym sense found represent the same semantic or not. After discarding the synonym senses that do not enrich the corresponding keyword sense, the result is a list of different possible senses for each keyword.

This process can be limited in time; obtaining the senses is executed in parallel for each keyword; within that task, the semantic enrichment of each keyword sense with its synonym senses is performed in parallel too.

[2] http://learn.tsinghua.edu.cn:8080/2003214945/travelontology.owl

3 Alignment of Senses

We explain in this section the sense alignment process, which is used in two situations by our system: 1) to check which synonym senses represent the same semantics as their keyword senses, and 2) to avoid redundancy in the list of possible senses of each user keyword. However both tasks share a common goal: to find when two given senses represent very similar semantics; in that case they will be considered synonyms and both senses will be integrated[3].

In order to decide if two senses must be integrated (as a single sense) or not, the system computes their *synonymy probability*. Thus the system avoid redundancy among the possible senses of a keyword. At present, several solutions to determine the matching among ontological terms of distinct ontologies have been proposed, see [9] for recent surveys. Our approach computes coefficients of synonymy degree in the [0,1] range, however other approaches as semantic matching [3] can be used as well.

The synonymy measure used relies on both *linguistic and structural* characteristics of ontologies. The following steps are performed: 1) an initial computation using linguistic similarity, which consider labels as strings, 2) a recursive computation using structural similarity, which exploits the semantics of terms (ontological context) until a certain depth, and 3) the above values are combined to obtain the resultant synonymy measure.

Our proposal for sense alignment is not just a comparison between two senses but an iterative process, which improves the quality and efficiency of ontology matching and enables the reuse of new discovered senses. In other words, each new integrated sense must be considered as candidate to integrate with the rest. For the same reason, new senses that do not integrate are stored because they could become the missing semantic gap between two senses. Although this method is costly (we limit its execution time), it performs a much better ontology alignment among senses. Due to space limitations, we do not detail this process.

In a variety of approaches, the similarity measure is only calculated among ontological terms that plays as classes. However, unlike another works, we propose a way to obtain the synonymy probability according to the type of senses that we compare. Details about this process is available in [1] as it not the main goal of this paper.

4 Conclusions

In this paper we have presented a semantics-guided approach to discover the possible senses for a set of keywords, by searching and extracting relevant knowledge from different ontology pools; ontology matching and synonymy estimation techniques are used to merge senses considered similar enough. The main features of our proposal are the following:

[3] The integration process that we propose can be found in [1].

1. It uses an iterative approach to retrieve from different knowledge repositories the possible senses of each user keyword, in a parallel manner. A sense is represented basically as the (multi)ontological context of a term, and the system is able to deal with senses corresponding to different kind of ontology terms (classes, properties, and individuals).
2. It considers not only the senses corresponding to ontology terms syntactically matching the user keywords but also the senses of ontology terms matching the synonyms of the user keywords, recursively, in order to semantically enrich the keyword senses retrieved within a certain synonymy threshold.
3. It measures the synonymy degree between two senses by considering their linguistic and structural similarity. Statistical techniques like sampling and parallel processing are used to improve the performance of this process.

We believe that this technique to find out the semantic different between senses (subsets of ontologies) can be applied to many fields. As example, we are currently working on using the retrieved senses to generate queries expressed in a knowledge representation language to retrieve data corresponding to the user keywords.

References

1. M. Espinoza, J. Gracia, R. Trillo, and E. Mena. Discovering the semantics of keywords: An ontology-based approach. In *The 2006 International Conference on Semantic Web and Web Services (SWWS'06), Las Vegas, Nevada (USA)*. CSREA Press, June 2006.
2. T. Finin, L. Ding, R. Pan, A. Joshi, P. Kolari, A. Java, and Y. Peng. Swoogle: Searching for knowledge on the Semantic Web. In *AAAI 05 (intelligent systems demo)*, July 2005.
3. F. Giunchiglia, P. Shvaiko, and M. Yatskevich. Discovering missing background knowledge in ontology matching. In *Technical Report DIT-06-005, Informatica e Telecomunicazioni, University of Trento*, February 2006.
4. J. Gracia, R. Trillo, M. Espinoza, and E. Mena. Querying the web: A multiontology disambiguation method. In *Sixth International Conference on Web Engineering (ICWE'06), Palo Alto (California, USA)*. ACM, July 2006.
5. T. R. Gruber. Towards principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*. Kluwer Academic Publishers, 1993.
6. B. Le, R. Dieng-Kuntz, and F. Gandon. On ontology matching problems - for building a corporate semantic web in a multi-communities organization. In *ICEIS (4)*, 2004.
7. V. Lopez, E. Motta, and V. Uren. Poweraqua: Fishing the semantic web. In *3rd European Sematic Web Conference, Budva, Montenegro*, June 2006.
8. G. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11), nov 1995.
9. P.Shvaiko and J. Euzenat. A survey of schema-based matching approaches. In *Journal on Data Semantics*, 2005.