Document indexing for automatic semantic annotation support

Marek Ciglan, Michal Laclavík, Martin Šeleng, Ladislav Hluchý
Institute of Informatics, Slovak Academy of Sciences, Dubravska cesta 9
845 07 Bratislava, Slovakia
marek.ciglan@savba.sk

Abstract

Nowadays, capturing the knowledge in ontological structures is one of the primary focuses of the knowledge management research. To exploit the knowledge from the vast quantity of existing unstructured texts available in natural languages in ontologies, tools for automatic semantic annotation (ASA) are heavily needed. In this paper, we present an approach to ASA and a method for documents content indexing and we describe how the rules mining from the indexed collection of documents can be beneficial for ASA. We use the statistical values obtained from indexed documents mining to estimate the accuracy of new instances of ontological concepts created by ASA. We evaluate this approach on a collection of domain specific documents and present preliminary results.
Keywords: automatic semantic annotation, information retrieval

1. Introduction

Adding machine processable information about documents content is one of main challenges of emerging semantic oriented systems. Ultimate goal is to allow machine based reasoning about content of vast quantity of documents produced by human experts and to allow automatic inference of new knowledge. One step towards this goal is to enable automatic and semi-automatic semantic annotation (ASA) of unstructured texts such as web pages, office documents, that provides the means to transfer useful information from the documents to the ontology structures.

In this paper, we present combination of a traditional method from information retrieval domain and an annotation method from semantic web, which increase the relevance of automatic annotation. Namely, we have integrated full-text indexing and search mechanism with ASA based on regular expression matching. Indexes of full-text search engine are exploited to gather statistical information about words occurrences in the document collection to estimate the relevance of the ASA outputs.

We present ASA and full-text indexing, including the summarization of the state-of-the-art and description of our approaches (Sections 2,3), we continue by description of the integration of Ontea annotation tool with indexing engine and present the experimental results. We conclude the paper with the summary and future work description.

2. Semantic Annotation

Automated annotation of the Web documents is a key challenge of the Semantic Web effort. Web documents are structured but their structure is understandable only for humans, which is the major problem of the Semantic Web. Annotation solutions can be divided into manual and semi-automatic methods. This different strategy depends on a use of the annotation. There is number of annotation tools and approaches such as CREAM [15]

or Magpie [16] which follow the idea to provide users with useful visual tools for manual annotation, web page navigation, reading semantic tags and browsing [18] or provide infrastructure and protocols for manual stamping documents with semantic tags such as Annotea[1], Rubby[2] or RDF annotation[3].
Semi-automatic solutions focus on creating semantic metadata for further computer processing, using semantic data in knowledge management [19] or in information extraction application. Semi-automatic approaches are based on natural language processing [11] [12], a document structure analysis [13] or learning requiring training sets or supervision [14]. Moreover, other pattern-based semi-automatic solutions such as PANKOW and C-PANKOW [10] exist, using also Google API for automatic annotation. Other methods exists which use variety of pattern matching mechanisms. Such solutions can be used also for annotating of media not only documents or text since results of media analysis can be set of patterns.

2.1 Ontea

One of pattern based solution is also Ontea [19] developed at IISAS. Ontea works on text, in particular domain described by domain ontology and uses regular expression patterns for semi-automatic semantic annotation. Ontea detects or create ontology elements/individuals within the existing application/domain ontology model according to defined patterns. Several cross application patterns are defined but in order to achieve good results, new patterns need to be defined for each application.

1 http://www.w3.org/2001/Annotea/
2 http://www.w3.org/TR/ruby/
3 http://ilrt.org/discovery/2001/04/annotations

3. Document indexing and search

Information retrieval is a process of identifying the text resources of interest from a large collection of documents that would satisfy the user needs. Full text search is a wildly used concept in today information systems for information retrieval. Full-text search engines usually operate over the index structure which keeps information about documents content. Indexes of the documents content are exploited because of the time efficiency of retrieving information from those structures.

3.1 Related work

Documents content indexing is a well established method for of information retrieval which crossed the border of academic research and become a part of every day's life. Full text search engines are used to find documents stored at users workstations (desktop search engines) as well as to locate resources in intranet and Internet. Main technological challenges addressed by document indexing and search solutions are: index Data Structures, performance (Maintenance, Lookup speed), transformation of words to their base form - usually done by stemming or lemmatization (this topic is discussed in more detail in section 3.2), provide rich query mechanism (phrase queries, wildcard queries, proximity queries, range queries), stop words filtering, search results ranking. Large number of number of document indexing solutions are available both under commercial and open source licenses with different level of features implementation. We mention several popular systems suitable for intranet and document repository indexing and searching: Apache Lucene [1] is a search engine designed for high-performance search, supporting large number of query mechanisms. Another search engine is OpenFTS [2] (Open Source Full Text Search engine) using relational database PostgreSQL as backend for storing indexes,

provides online indexing of data and relevance ranking. MnoGoSearch [3] is a search engine designed primary for indexing HTML content with HTML specific features such as META tags support, robots exclusion standard support.

3.2 Words base forms

One of the driving factors for developing yet another indexing and search engine was the study of stemming and lemmatization methods for Slovak language and subsequent integration of suitable methods with the search engine. Therefore, we describe in this subsection basic approaches to the words' base form acquisition. Different morphological variants of the natural languages words have in most cases the same or very similar semantic interpretations and can be considered as equivalent for the purpose of information retrieval systems. This means that different morphological forms can be represented by a single representative term. Queries can than produce more relevant responses and the dictionary size needed for representing a set of documents decrease. A smaller dictionary size results in a saving of storage space and processing time. Two main approaches to words' base form acquisition are lemmatization and stemming. Lemmatization uses the dictionaries produced by human experts to retrieve the base form of a given word. Wordnet [4, 5] is one sophisticated dictionaries for English language, that can be used for lemmatization. Stemming is a method, which algorithmically derives the stem of a given word; stems produced by stemming algorithms often do not belong to the given natural language, however they identify the class of words from natural language. Popular stemming algorithm for English language is Porter algorithm [8,9].

3.3 RFTS

We have developed a tool for document indexing and document search, named RFTS (Rich full-text search). The motivation for implementing another search engine was to have an easily extendable and configurable document indexing tool to evaluate novel methods for information retrieval, documents statistical analysis and lemmatization and stemming methods for Slovak language. Detailed information about documents content is stored in the index structure, including the positions of the word in the documents, phrase number within the document. The words in tool's dictionary are kept in the basic form, different stemming algorithms are used for documents in different languages. The tool exploit relational database to store all the information about documents and its contents. From engineering point of view, it is worth to mention that RFTS functionality in conjunction with Corporate Memory [7] (also developed within project NAZOU [6]) can be accessed locally (using JAVA interfaces or command line tools) as well as remotely using RPC calls or Web Service interface. The remote access and Web Service interface allows easy integration of the RFTS indexing and search solution in other components and allows rapid prototyping of new tools that require full-text search or some form of statistical analysis of document collection.

4. Ontea supported by RFTS

Ontea's method of automatic annotation based on regular expressions matching showed promising results for domain specific texts. However it suffers from frequent mismatching which leads to creation of imprecise instances of ontological concepts. We propose to overcome this obstacle by evaluating the relevance of candidate instances by the means of statistical analysis of the occurrence of the matched words in the document collection. Based on regular expression, Ontea identifies part of a text related to semantic context and match the subsequent sequence of characters to create an instance of the concept. Let us denote the sequence of words related to semantic

context by C and word sequence identified as a candidate instance as I. We evaluate the relevance of the new instance by computing the ration of the close occurrence of C and I and occurrence of I: *close_occurrence(C, I) / occurrence(I)*
RFST indexing tool provides us with enough functionality to retrieve required statistical values computed from the whole collection of documents stored in RFTS index structures.
Let *COLL* be a collection of the documents $d_1,..,d_n$: $COLL = d_{1,..}, d_n$

Let d *in COLL,* $distance \in \mathbb{N}$, and $w_{1,..}, w_k$ are the words from natural language.
Function $dist(d, distance, w_1, w_{2,..}, w_k)$, where $k \leqslant distance$, denotes the number of distinct word sequences of the length *distance* containing the words $w_{1,..}, w_k$.
we compute the relevance of candidate instance as:

$$relevance(C, I, wordsdist) = \frac{\sum dist(d, wordsdist, C \cup I)}{\sum dist(d, (I), I)}$$

If the resulting relevance value exceeds defined threshold, the candidate word sequence *I* is considered to be a valid instance of the semantic concept related to sequence *C*. For the experimental evaluation of the approach, the threshold was set manually after inspecting the preliminary relevance values of the generated candidate instances. The utilization of RFTS brings also an important benefit of treating different morphological word forms as a single class of equivalence represented by the word stem or lemma. All the documents that are subject to semantic annotation by Ontea must be part of the document collection indexed by RFTS tool.

5. Evaluation

In this chapter we discuss the algorithm evaluation and success rate.

To evaluate the performance of annotation, we used the standard recall, precision and F_1 measures. Recall is defined as the ratio of correct positive predictions made by the system and the total number of positive examples. Precision is defined as the ratio of correct positive predictions made by the system and the total number of positive predictions made by the system:

$$Recall = \frac{Match}{Count} = \frac{Relevant\ retrieved}{All\ relevant} \quad (1)$$

$$Precision = \frac{Match}{Ontea} = \frac{Relevant\ retrieved}{All\ retrieved}$$

Recall and precision measures reflect the different aspects of annotation performance. Usually, if one of the two measures is increasing, the other will decrease. These measures were first used to measure IR (Information retrieval) system by Cleverdon [11]. To obtain a better measure to describe performance, we use the F1 measure (first introduced by van Rijsbergen [12]) which combines precision and recall measures, with equal importance, into a single parameter for optimization. F1 measure is weighted average of the precision and recall measures and is defined as follows:

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (2)$$

5.1 Test set of documents

As reference test data, we used 500 job offers downloaded from web using wrapper which prepared us some structured data. This was converted to a defined ontology and manually checked and edited according to 500 html documents representing reference job offers. Ontea processed reference html documents using the reference ontology resulting in new ontology metadata consisting of 500 job offers, which were automatically compared with reference manually checked job offers ontology metadata.

5.2 Target ontological concepts for identification

In this test, Ontea used simple regular expressions matching from 1 to 4 words starting with a capital letter. This is experiment is referred as "Ontea" in next chapter. In second case we used domain specific regular expressions which identified locations and company names in text of job offers and Ontea was also creating individuals in knowledge base, while in the first case and Ontea did not create extra new property individuals only searched for relevant individuals in knowledge base. This second case is referred as "Ontea creation". Third case used also previously described RTFS indexing tool to find out if it is feasible to create new individual using word occurrence functionality of RTFS this case is referred as "Ontea creation, RFTS"
So we did our experiments in 3 cases:

- Ontea: searching relevant concepts in knowledge base (KB) according to generic patterns
- Ontea creation: creating new individuals of concrete application specific objects found in text
- Ontea creation, RTFS: Similar as previous with the feedback of RTFS to get relevance computed above word occurrence. Individuals were created only when relevance was above defined threshold which was set up to 10%

We have used following regular expressions:

- Generic expression matching one or more words in text. This was used only to search concepts in KB ([A-Z][-A-Za-z0-9]+[\\s]+ [-a-zA-Z]+)
- Identifying geographical location in text and if not found in KB individual was created
  Location:[\\s]*([A-Z][-a-zA-Z]+[ ]*[A-Za-z0-9]*)
- Identifying company in the text, this was used also with other abbreviations such as "Ltd" or "a.s.", "s.r.o." for Slovak language
  [\\s]+([-A-Za-z0-9][ ]*[A-Za-z0-9]*),[ ]*Inc[\\.\\s]+

5.3 Experimental results

Experimental results using precession, recall and F1-measures are in table 1. In the table we compare our results with other semantic annotation approaches and we also list some advantages and disadvantages. The column "relevance" is in case of Ontea F1-measures but in case of other methods it can be evaluated by other techniques and usually it is not common. For example for C-PANKOW, relevance is referred as recall.

The experimental results are summarized in table 1. Rows relevant to our annotation approach are in grey color, were we are showing success rate of three evaluation cases mentioned in previous chapter. The row "Ontea creation, RTFS" case is most important concerning evaluation were we combined indexing and annotation techniques. By using this combination we were able to eliminate some not correctly annotated results. For example by using "[Cc]ompany[:\\s]*([A-Z][-A-Za-z0-9][ ]*[A-Za-z0-9]*" regular expression in second case we have created and identified companies such as "This position" or "International company" which were identified as not relevant in third case with RTFS.

Similarly "Ontea creation" identified also companies like Microsoft or Oracle which is correct and in combination with RTFS this was eliminated. Because of this issue recall is decreasing while precession is increasing. Here it seems that RTFS case is not successful but opposite is true because in many texts Microsoft is identified as products e.g. "Microsoft Office" so if we take more text to annotate it is better to not annotate Microsoft as company and decrease recall. If we would annotate Microsoft as company in other texts, used in context of "Microsoft Office" we will decrease precision of annotation. So it is very powerful to use presented annotation technique in combination with indexing in applications were precession need to be high.

6.Conclusion

By integrating information retrieval system based on full-text indexing and search (RFTS) and semantic annotation tool (Ontea), we were able to improve the results of the automatic semantic annotation process for domain specific documents – increasing the precision of newly created instances at least by 4%. However, the recall of identified instances decreased. This is an advantageous trade-off as the ontological data precision is the primary goal of our work on automatic annotation.

We have also identified, but not proved yet, that using large collection of experimental texts or documents "Ontea creation" (without RTFS indexing) precession will decrease and in combination with RTFS precision will still stay over 90%, which is very high for semi-automatic annotation solution.

Ontea algorithm disadvantage is requirement to set up domain specific patterns. While annotation methods as C-PANKOW are more generic, Ontea is a simpler, faster solution with a better success rate, suitable for knowledge management, information extraction or knowledge acquisition applications, where large number of documents needs to be annotated.

Main advantages of described method are: supporting of Slovak language, fast algorithm comparing to other methods, instance duplicity identification and very high precision.

	Method	relevance %	precision %	recall %	disadvantages	Adventages
Ontea	regular expresions, search in knowledge base (KB)	**71**	64	83	high recall, lower precesion	high succes rate, generic solution, solved duplicity problem, fast algorithm
SemTag	disambiguatiy check, searching in KB	high	high		works only for TAP KB and English	fast and generic solution
Ontea creation	regular expresions (RE), creation of individuals in KB	**83**	**90**	76	aplication specific patterns are needed	support Slovak language
Ontea creation RTFS, TS	RE, creation of individuals in KB + RFTS	73	**94**	69	low recall	disambiguities are found and not annotated
Wrapper	document structure	high	high		zero success with unknown structure	high success with known structure
PANKOW	pattern matching	59			low success rate	generic solution
C-PANKOW	POS taging and pattern matching Qtag library	74		74	suitable only for English, slow algorithm	generic solution
Hahn et al.	semantic and syntactic analysis	76			works only for English not Slovak	
Evans	clustering	41			low success rate	
Human	manual annotation	high	high	high	problem with creation of individuals duplicities, inacuracy	high recall and precesion

Table 1: Experimental results

7.Future work

Presented work is an intermediate result on our research on automatic and semi-automatic semantic annotation. Subsequent effort will be focused on studying and tuning instance relevance computation from the document collection. We plan to study the effect of increasing the distance parameter (distance larger than cardinality of C and I sequences), relevance computation based on C and I membership in a phrase in a documents (instead of word distance concept), document preprocessing methods that would pre-format the selected terms in order to increase regular expressions matching precision. We will study the effects of extending the document collection (that form the basis for our

statistical analysis) by text, which do not belong to the specific domain and we will examine the results obtained from document from different domains. We will also analyze the effects of document collection size on the instance relevance identification.

Acknowledgments

This work is partially supported by projects NAZOU SPVV 1025/2004 and VEGA 2/7098/27.

References

[1] Hatcher E., Gospodnetić O., Lucene in Action, Manning (12 Jan 2005), ISBN: 1932394281

[2] OpenFTS- http://openfts.sourceforge.net

[3] mnoGoSearch - http://www.mnogosearch.org

[4] Miller, G.:WordNet: An On-line Lexical Database, Special Issue, International Journal of Lexicography, Vol. 3, Num. 4, 1990

[5] WordNet: http://wordnet.princeton.edu/

[6] Návrat, P., Bieliková, M., Rozinajová, V.: Methods and Tools for Acquiring and Presenting Information and Knowledge in the Web. In: CompSysTech 2005, B. Rachev, A. Smrikarov (Eds.), Varna, Bulgaria, June 2005. pp. IIIB.7.1-IIIB.7.6.

[7] M. Ciglan, M. Babik, M. Laclavik, I. Budinska, and L. Hluchy. Corporate memory: A framework for supporting tools for acquisition, organization and maintenance of information and knowledge. In J. Zendulka, editor, Proc. of 9th Int. Conf. on Information Systems Implementation and Modelling (ISIM 2006), pages 185--192. MARQ, Ostrava, 2006.

[8] Jones, K. S., Willet, P: Readings in Information Retrieval, San Francisco: Morgan Kaufmann,1997, ISBN 1-55860-454-4.

[9] Van Rijsbergen, C. J., Robertson, S. E., Porter, M. F. New models in probabilistic information retrieval. British Library Research and Development Report, no. 5587, British Library, London, 1980.

[10] Cimiano P., Ladwig G., Staab S.: Gimme' the context: context-driven automatic semantic annotation with c-pankow. In WWW '05, pages 332-341, NY, USA, 2005. ACM Press. ISBN 1-59593-046-9.

[11] Madche A., Staab S.: Ontology learning for the semantic web. IEEE Intelligent Syst., 16(2):72-79, 2001

[12] Charniak E., Berland M.: Finding parts in very large corpora. In Proceedings of the 37th Annual Meeting of the ACL, pages 57-64, 1999.

[13] Glover E., Tsioutsiouliklis K., Lawrence S., Pennock D., Flake G.: Using web structure for classifying and describing web pages. In Proc. of the 11th WWW Conference, pages 562-569. ACM Press, 2002.

[14] Reeve L., Hyoil Han: Survey of semantic annotation platforms. In SAC '05, pages 1634-1638, NY, USA, 2005. ACM Press. ISBN 1-58113-964-0. doi: doi.acm.org/10.1145/1066677.1067049

[15] Handschuh S., Staab S.: Authoring and annotation of web pages in cream. In WWW '02, pages 462-473, NY, USA, 2002. ACM Press. ISBN 1-58113-449-5. doi: http://doi.acm.org/10.1145/511446.511506.

[16] Domingue J., Dzbor M.: Magpie: supporting browsing and navigation on the semantic web. In IUI '04, pages 191-197, New York, NY, USA, 2004. ACM Press. ISBN 1-58113-815-6.

[17] Uren V. et al.: Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Journal of Web Semantics: Science, Services and Agents on the WWW, 4(1):14-28, 2005.

[18] Uren V. et al.: Browsing for information by highlighting automatically generated annotations: a user study and evaluation. In K-CAP '05, pages 75-82, NY, USA, 2005b. ACM Press. ISBN 1-59593-163-5

[19] Michal Laclavik, Martin Seleng, Emil Gatial, Zoltan Balogh, Ladislav Hluchy: Ontology based Text Annotation – OnTeA; Information Modelling and Knowledge Bases XVIII. IOS Press, Amsterdam, Marie Duzi, Hannu Jaakkola, Yasushi Kiyoki, Hannu Kangassalo (Eds.), Frontiers in Artificial Intelligence and Applications, Vol. 154, February 2007, pp.311-315. ISBN 978-1-58603-710-9, ISSN 0922-6389.