ONTOLOGY-DRIVEN CONCEPTUAL DOCUMENT CLASSIFICATION

Gordana Pavlović-Lažetić and Jelena Graovac
Faculty of Mathematics, University of Belgrade, Studentski trg 16, Belgrade, Serbia

Keywords: Document classification, Wordnet, SWN, Ontology, Proper name.

Abstract: Document classification based on the lexical-semantic network, wordnet, is presented. Two types of document classification in Serbian have been experimented with – classification based on chosen concepts from Serbian WordNet (SWN) and proper names-based classification. Conceptual document classification criteria are constructed from hierarchies rooted in a set of chosen concepts (first case) or in hierarchies rooted in some of the proper names' hypernyms (second case). A classificator of the first type is trained and then tested on an indexed and already classified Ebart corpus of Serbian newspapers (476917 articles). Precision, recall and F-measure show that this type of classification is promising although incomplete due mainly to SWN incompleteness. In the context of proper names-based classification, a proper names ontology based on the SWN is presented in the paper. A distance based similarity measure is defined, based on Euclidean and Manhattan distances. Classification of a subset of Contemporary Serbian Language Corpus is presented.

1 INTRODUCTION

Different document processing tasks such as document retrieval, internet web pages search, information extraction an the like, may highly benefit from conceptual document classification. For example, search for the term Moon will be more efficient if conducted in a class corresponding to celestial bodies, than in the whole set of documents, thus increasing the precision of the search. On the other side, in order to search for celestial bodies, it is reasonable to expand the conceptual hierarchy and to search for Moon, thus increasing the recall.

There are different pre-specified classification schemes, e.g., EAGLES Guidelines (EAGLES, 1996), Library of Congress Classification (LCC, 2009), Reuters news archive (Reuters, 2010). Ebart (Ebart, 2010) is the largest digital media documentation in Serbia, with more than 1.200.000 completely indexed texts of fifteen complete daily and weekly press, since 2003. It is classified in a usual way into internal politics, foreign politics, society, economics, chronicle and crime, culture and entertainment, sport, media, etc.

This paper proposes a document classification in Serbian based on wordnet and distance based algorithms. Wordnet is a semantic lexical resource addressing conceptual hierarchies issue (Miller, 1995) and is under development for Serbian (SWN - Serbian WordNet, (Krstev et al., 2004)). Wordnet is a de facto standard for semantic networks, and it has been used so far in document classification as a source of synonyms mainly (Scott and Matwin, 1998), (Rosso et al., 2004), (Rodriguez et al., 1996).

As the first experiment, classification has been applied to a part of Ebart corpus (articles from the Serbian daily newspaper "Politika"). Each class is defined so that a set of concepts from the wordnet, as well as all of their hyponyms, have been assigned to it. An article is assigned to a class for which it contains the largest number of concepts (with repetitions) assigned to that class. The results have been validated using statistical measures of precision, recall and their combination - F-measure.

As the second experiment, the classification based on ontology of proper names for a predefined set of classes has been applied to a subset of the corpus of contemporary Serbian language developed at the Faculty of mathematics, University of Belgrade. Distance based similarity measures are defined that assign documents to classes on the lowest distance, taking into account ratio of the number of occurrences of proper names from the corresponding conceptual hierarchy in the document and in the whole collection.

Pavlović-Lažetić G. and Graovac J..
ONTOLOGY-DRIVEN CONCEPTUAL DOCUMENT CLASSIFICATION.
DOI: 10.5220/0003063903830386
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 383-386
ISBN: 978-989-8425-28-7

2 LEXICAL RESOURCES FOR CONCEPTUAL DOCUMENT CLASSIFICATION IN SERBIAN

Lexical resources for Serbian have been developed within the Human Language Technologies Group at the Faculty of Mathematics, University of Belgrade (HLTG, 2010). Except for Serbian language corpus and multilingual parallel corpora, especially important are the system of electronic morphological dictionaries of Serbian and the lexical-semantic network, Serbian WordNet, SWN (Vitas et al., 2003).

The Serbian WordNet has been initiated in the scope of BalkaNet (BWN), the Balkan wordnet project (2002-2004) aimed at producing a multilingual database with wordnets for five Balkan languages (Greek, Turkish, Bulgarian, Romanian and Serbian) as well as Czech (Tufis et al., 2004) . BWN is based on the model of the EuroWordNet (EWN), a multilingual database with wordnets for Dutch, Italian, Spanish, German, French, Czech and Estonian (Vossen, 1998). The structure of these wordnets is basically the same as the structure of the Princeton WordNet (PWN) for English (Fellbaum, 1998). Wordnet consists of synsets - the sets of synonymous words representing a concept, with basic semantic relations between them forming a semantic network.

At the moment, SWN consists of around 15000 synsets. There are 9 noun hierarchies (top ontologies) rooted at generic concepts of ”abstraction” ”act”, ”event”, ”entity”, ”phenomenon”, ”group”, ”psychological feature”, ”state” and ”possession”. Hierarchies have different depths (up to 12 levels). Somewhere in the middle of these hierarchies concepts are found which are neither too general, nor too specific. Those concepts are called ”basic concepts” and they are used for classification.

Figure 1 represents conceptual hierarchy having the proper name ”Zeus” at its leaf, as well as all the other proper names that are hyponyms of the Zeus’s hypernym ”Greek deity” (other Greek deities). If ”Greek deity” is used as the basic concept describing a document class, than all the proper names – names of Greek deities, will be part of the class description.

3 DOCUMENT CLASSIFICATION BASED ON CHOSEN CONCEPTS

The first experiment has been performed on a subset of Ebart corpus consisting of articles in the following columns: sport, economics, politics, culture and entertainment, chronicle and crime, published from 2003 to 2006. There are 476917 such articles. The classification process proceeds as follows:

1. the chosen columns (article types) are assigned to predefined set of classes (we experimented with 3, 4 and 5 classes);
2. key words for each column and each class are identified as the most frequent words in a set of articles from the given column / class (training set);
3. SWN concepts containing the chosen key words, along with all of their hyponyms, are assigned to the corresponding classes;
4. class assignment functions are defined for an article (from the test set) in different ways, the simplest being the maximum number of occurrences of literals from the hierarchy rooted in the concepts assigned to the class, maybe filtered by domains.

The chosen columns have been assigned the following SWN concepts and key words (translated in English):

SPORT

event → social event → **contest, competition**
group → social group → **team**
event . . . →**result, victory, triumph, defeat**
act → activity → recreation → **sport**

and similarly for the classes ECONOMICS, POLITICS, CULTURE AND ENTERTAINMENT and CHRONICLE AND CRIME.

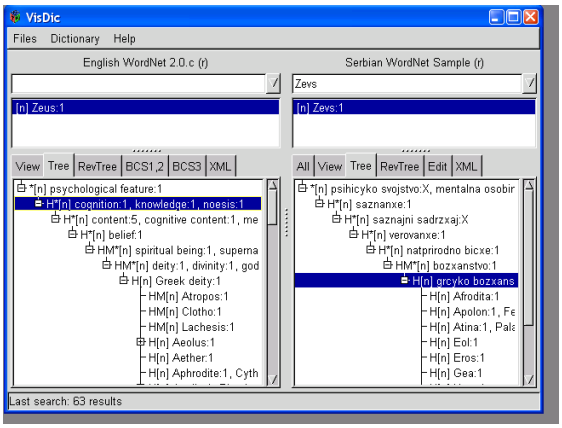

Figure 1: Proper names that are in the same conceptual hierarchy (Greek deity) as the given proper name (Zeus).

3.1 Results

The table 1 illustrates the results of the experiment performed - measures of precision, recall and F-measure for classification into three classes (similarly into four and five classes).

Table 1: Results of 3-classification.

class	*precision*	*recall*	*F-measure*
sport	97.30%	87.21%	91.98%
politics	91.27%	76.20%	83.06%
economics	59.19%	93.57%	72.51%

It may be noticed that the source of texts is not in favor of this approach, since daily press is characterized by concise and factual expression. The approach is expected to produce even better results on rich vocabulary corpora.

Instead of chosen concepts, document classification may also be based on the most productive concepts (Tomašević and Pavlović-Lažetić, 2008) in each of the wordnet top ontology hierarchies.

4 PROPER NAME ONTOLOGY-BASED CLASSIFICATION

Since document retrieval is usually performed on proper names, document classification may be based on an ontology of proper names. Proper names for document classification in Serbian may be extracted from the SWN.

Classification criteria may be constructed from the hierarchy rooted in the most productive concept which is in hypernym relation with the proper name, the so-called proper name's ontology term. For example, for the proper name "Zemlja" (Eng. Earth), third-level hypernym, "nebesko telo" (Eng. celestial body), may be used as a proper name ontology term, and the hierarchy, rooted at "nebesko telo", would constitute the basis for definition of the corresponding class. Out of all the SWN synsets, about 10% contain proper names. Ontology terms are found at different levels of proper name hypernymy (usually the third or fourth).

Departing from proper names in the present SWN noun hierarchies, we have come up with the proper names ontology schema consisting of 23 proper name ontology terms, e.g., entity/person, entity/inhabitant, entity/city, entity/celestial body, group/dinasty, group/organization, etc.

4.1 Classification Method

Classification of documents based on proper names ontology, is performed using distance-based approach. There are as many classes as there are proper name ontology hierarchies considered, at the moment it is around 20. Each class $C_i \in C$ is defined by a tuple t_i=($C_{i1}, C_{i2}, \ldots, C_{ii_k}$), where C_{ij} represents a proper name from the corresponding hierarchy. The tuple t_i is a representative of the class C_i. Given a database $D = \{D_1, D_2, \ldots, D_n\}$ of documents, each document $D_i \in D$ is assigned to a class C_j such that dist(D_i, C_j) $\leq$ dist(D_i, C_h) for each $C_h \in C, C_h \neq C_j$.

Distance may be defined by different formulas for similarity (dissimilarity) between database documents or between a database document and a class. It also depends on how numeric characteristics of similarity on a single attribute (proper name) is defined. We considered (and experimented with) variations of two of document distance formulas, Euclidean and Manhattan distance measures (Tan et al., 2006).

4.2 Results

The classification based on ontology of proper names has been applied to a subset of the corpus of contemporary Serbian language developed at the Faculty of mathematics, University of Belgrade, and accessible on the web at http://korpus.matf.bg.ac.rs. The collection experimented with consists of 234 non-tagged texts, total size of around 23 million words, and encompassing texts from the newspapers and journals ("Politika", "Journal of the Serbian Orthodox Church", "Viva", "Svet", "NIN", "Magazin", "Ilustrovana politika"), fiction, textbooks, chapters from literature pieces, etc.

For experimental purposes, we chose five conceptual hierarchies from the SWN rooted at proper names ontology terms "država" (Eng. country), "dinastija" (Eng. dynasty), "nebesko telo" (Eng. celestial body), "manastir" (Eng. monastery) and "hrišćanski praznik" (Eng. Christian holyday).

Since the SWN itself is in early development phase, the corresponding conceptual hierarchies contain limited number of proper names, between 3 (for "dinastija") and 16 (for "država"). The classification is performed following the procedure for crisp classification described in the previous section. There were 195 documents containing at least one occurrence of at least one proper name from at least one of the chosen conceptual hierarchies. Although similarity measures were somewhat different, classification obtained by both Euclidean and Manhattan distance formulas were identical.

Assignment of documents to classes is quite successful. For example, among the 59 documents classified as belonging to the class "nebesko telo", the top most 10 belong to geography textbooks, science fiction books and the newspaper texts dealing with the subject. Among the documents classified as belong-

ing to the class ”država”, top most 10 belong to daily and weekly newspapers (”Politika” and ”Nin”).

5 CONCLUSIONS

Document classification may benefit from both common and proper names. Proper names are especially suitable when applied for subclassification of documents already classified into common classes such as news, articles, science, politics, finance, etc. Two types of classification based on lexical-semantic network wordnet are presented. Classification based on wordnet basic concepts (common words) proved quite successful in classifying a large newspaper archive into several classes. A distance-based classification driven by an ontology of proper names is then presented, where conceptual hierarchies of proper names follow the structure of the Serbian WordNet. Results of classification applied to an untagged part of the corpus of contemporary Serbian, of around 23 million words, are presented.

Our future plans involve improvement of the proposed classification framework and its implementation – enlargement of SWN and lookup at EWN, as well as development of new classification methods based on character and word patterns in texts. We also plan to define new distance measures and to perform a multi-way comparison of different classification methods applied to different types of corpora.

ACKNOWLEDGEMENTS

The work presented has been financially supported by the Ministry of Science and Technological Development of the Republic of Serbia, Project No.148921A.

REFERENCES

EAGLES (1996). *Preliminary Recommendations on Text Typology, EAGLES Document EAG-TCWG-TTYP/P*. Expert Advisory Group on Language Engineering Standards, European Commission.

Ebart (2010). *Aktuelna arhiva*. Medijska dokumentacija Ebart, http://www.arhiv.rs.

Fellbaum, C. (1998). *Wordnet: An Electronic Lexical Database*. The MIT Press.

HLTG (2010). *Resursi srpskog jezika*. Human Language Technologies Group, http://korpus.matf.bg.ac.rs, Faculty of Mathematics, University of Belgrade.

Krstev, C., Pavlović-Lažetić, G., Vitas, D., and Obradović, I. (2004). Using textual and lexical resources in developing serbian wordnet. In *Romanian J. Sci. Tech. Inform. (Special Issue on Balkanet), 7(1-2)*, pages 147–161. Romanian Academy.

LCC (2009). *Library of Congress Classification Outline*. http://www.loc.gov/catdir/cpso/lcco/, U.S. government.

Miller, G. (1995). Wordnet: A lexical database. In *Comm. ACM 38(11) 39–41*. ACM – Association for Computing Machinery.

Reuters (2010). *Site Archive*. Thomson Reuters Corporate, http://in.reuters.com/resources/archive/in/index.html.

Rodriguez, M., Gomez-Hidalgo, J., and Diaz-Agudo, B. (1996). Using wordnet to complement training information in text categorization. In *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access, Bulgaria*.

Rosso, P., Molina, A., Pla, F., Jiménez, D., and Vidal, V. (2004). Text categorization and information retrieval usingwordnet senses. In *CICLing 2004, Lecture Notes in Computer Science, 2945.*, pages 596–600. Springer- Verlag.

Scott, S. and Matwin, S. (1998). Text classification using wordnet hypernyms. In *Usage of WordNet in Natural Language Processing Systems1st International Wordnet Conference*. http://www.ceid.upatras.gr/Balkanet/files/balkanet-elsnet-ko-accept.pdf.

Tan, P., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*. Addison-Wesley.

Tomašević, J. and Pavlović-Lažetić, G. (2008). Productivity of concepts in serbian wordnet. In *Proceedings of the Sixth Language Technologies Conference: proceedings of the 11th International Multiconference Information Society - IS 2008, 86–91*, pages 86–91.

Tufis, D., Cristea, D., and Stamou, S. (2004). Balkanet: Aims, methods, results and perspectives. a general overview. In *Romanian J. Sci. Tech. Inform. (Special Issue on Balkanet), 7(1-2), . 9–43*, pages 9–43. Romanian Academy.

Vitas, D., Pavlović-Lažetić, G., Krstev, C., Popović, L., and Obradović, I. (2003). Processing serbianwritten texts: An overview of resources and basic tools. In *Proceedings of the International Workshop on Balkan Language Resources and Tools, Thessaloniki*, pages 97–104.

Vossen, P. (1998). *EuroWordnet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.