UNIVERSITY
OF TRENTO

DIPARTIMENTO DI INGEGNERIA E SCIENZA DELL'INFORMAZIONE

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.disi.unitn.it

LIGHTWEIGHT PARSING OF NATURAL LANGUAGE METADATA

Aliaksandr Autayeu, Fausto Giunchiglia, Pierre Andrews, Ju Qi

May 2009

Technical Report # DISI-09-028

Also: in Natural Language Processing for Digital Libraries (NLP4DL) Workshop, Viareggio, Italy, June 15th 2009.

Lightweight Parsing of Natural Language Metadata

Aliaksandr Autayeu, Fausto Giunchiglia, Pierre Andrews, Ju Qi

Department of Information Engineering and Computer Science, University of Trento[1]
{autayeu,fausto,andrews,qi}@disi.unitn.it

Abstract. Understanding metadata written in natural language is a premise to successful automated integration of large scale language-rich datasets, such as digital libraries. In this paper we describe an analysis of the part of speech structure of two different datasets of metadata, show how this structure can be used to detect structural patterns that can be parsed by lightweight grammars with an accuracy ranging from 95.3% to 99.8%. This allows deeper understanding of metadata semantics, important for such tasks as translating classifications into lightweight ontologies for use in semantic matching.

Introduction

Development of information technologies turned the data drought into the data deluge. This seriously complicated data management problems and increased the importance of metadata.

The amount of existing attempts (see surveys [1, 2]) to solve the semantic heterogeneity problem shows its importance and reveals the variety of domains where it applies. The state of the art algorithms try to apply to generic problem definitions at the schema level [7] and their large-scale evaluations [8] show two important directions for improvement: a) increasing the background knowledge [9] and b) improving natural language understanding [5].

Digital libraries metadata extensively use natural language, both in structured and unstructured form. Parsing of natural language metadata into a machine tractable language can enable a better understanding of its content, consequently improving semantic matching and integration algorithms. Natural language metadata is a novel domain of language understanding, and the current language processing technologies need a domain adaptation [10] to fit new domains with specific constraints such as metadata language structure. Moreover, the size of the current datasets [8] (see the following section) poses additional requirements on processing speed.

In this paper we analyse the language used in the metadata provided by the Library of Congress Subject Headings and the Open Directory Project. We show that the natural language used in these metadata is simple and can be accurately parsed by lightweight grammars. Parsers based on these grammars will allow a deeper understanding of the metadata semantics without sacrificing performance. We use these parsers for translating classifications into propositional description logics for use in semantic matching.

[1] This work has been partly supported by the INSEMTIVES project (FP7-231181, see http://www.insemtives.eu).

Datasets of Metadata

We have chosen the Library of Congress Subject Headings[2] (LCSH) and the Open Directory Project[3] (DMOZ) as representative datasets of language-rich metadata. In this section we explain our choice and provide some figures describing these datasets.

Libraries across the world use *subject headings* to classify books by subjects. Each library may create its own set of *subject headings* and thus will contribute to the heterogeneity of different classifications for the same subjects in this domain. The Library of Congress Subject Headings is a set of *subject headings*, maintained by the United States Library of Congress. It is the largest and most widespread set of *subject headings*, created by trained librarians and written in *natural language*. It covers a large number of topics and thus is not focused on a specific domain of human knowledge. All these reasons make the LCSH an interesting and representative sample of bibliographic metadata.

The Open Directory Project is a large scale on-line catalogue of web sites. Essentially, it contains categories classifying links to web sites and their description. Our primary interest lies in the DMOZ *category names* as they serve a similar purpose as the *subject headings* of the Library of Congress. However, in contrast with the latter, DMOZ is edited collectively by non-professional contributors who also use *natural language* to create category names. DMOZ covers various topics and is not focused on a specific domain. These properties make DMOZ *category names* an interesting and representative sample of catalogue metadata.

In the presented analysis, we use the hierarchical representation of both datasets where we select, for practical reasons, a random subset that is manually annotated with part of speech tags using the PennTreeBank tag set [3]. The annotation was performed by a single expert annotator. **Table 1** provides details of the dataset characteristics.

Characteristic	**LCSH** *(in headings)*	**DMOZ** *(in categories)*
Dataset size	335856	494040
Annotated subset size	44490	27976

Table 1. Dataset characteristics

While DMOZ dataset contains *category names* and LCSH dataset contains *subject headings*, both kinds of metadata are *labels* used for classification purposes and below we will use the term label when we mean both of them.

Part Of Speech Tagging

Parts of speech (POS) tags provide a significant amount of information about the language structure. The tagging is a fundamental step in language processing tasks such as parsing, clustering or classification. This is why we start our analysis with a look at POS tags of our datasets.

[2] http://authorities.loc.gov/
[3] http://dmoz.org

We use the OpenNLP toolkit[4] to automatically annotate the full datasets. First, using the manually annotated subset of each datasets, we test the performance of the standard OpenNLP tokenization and tagging models, which are trained on the Wall Street Journal and Brown corpus data. We then train our own tokenization and tagging models and analyse their performances. For the annotation analysis presented in the next section, we use the best performing models. To indicate the percentage of correctly processed tokens and labels (i.e. headings or categories) we report the precision per token (PPT), and the precision per label (PPL).

The standard OpenNLP tagging model's[5] performance is reported in the last row of **Table 2** as a baseline for both datasets. We report in bold the best performances of our custom trained tagging model (using a 10-fold cross-validation). To evaluate how well the models perform on the other dataset, we test the DMOZ model on LCSH dataset and vice-versa. The LCSH model is the best performer across all datasets while the DMOZ model performs poorly on the LCSH dataset. We suppose that the major reasons for this difference in performances are the lack of context in labels, different capitalization rules, different use of commas and various degree of representation of these phenomena in each dataset.

	DMOZ		LCSH	
	PPT, %	PPL, %	PPT, %	PPL, %
DMOZ	**95.03**	**93.74**	40.83	20.40
LCSH	86.61	82.05	**96.44**	**89.88**
OpenNLP	64.47	49.88	72.62	27.17

Table 2. Part of speech tagger performance

Language Structure Analysis

The training of the part of speech (POS) tagger reported in the previous section enabled the study of the language structure of the labels in the datasets. From this study, we can identify structural patterns in the language used in the subject headings of the LCSH and the categories of DMOZ. From these patterns, we can derive simple grammars that are discussed in the next section.

The first observation from the DMOZ dataset is that labels tend to be short phrases. In fact, more than half (50.83%) of all DMOZ category names contain only one token. 2- and 3-token names represent 17.48% and 27.61%, respectively, while the longer labels only occupy fewer than 5%. In comparison, the subject headings in LCSH tend to be longer and more complex, with only 8.39% of them containing one token, 20.16% two tokens and about 10-14% for each of 3-, 4-, 5- and 6-token headings; the remaining 11.45% of headings contains more than 6 tokens.

The DMOZ category names show a clear separation between "common" and "proper" category names. Namely, 37.11% of all category names are proper names. 99.43% of words

[4] http://opennlp.sourceforge.net/

[5] Trained on Wall Street Journal and Brown corpus data, which is reported to achieves >96% accuracy on unseen textual data.

used in remaining "common" category names (62.89% of all category names) are tagged with the following four POS tags: nouns (NN, 51.59%), plural nouns (NNS, 22.96%), conjunctions (CC, 18.50%) and adjectives (JJ, 6.38%). The remaining tokens are tagged with comma (,), preposition (IN), cardinal (CD), possessive (POS), singular and plural proper nouns (NNP and NNPS), comparative adjectives (JJR) and "to" (TO). That means that only 12 out of the whole set of 36 POS tags are actually used in the whole of DMOZ.

In comparison, the LCSH headings show no clear separation between "common" and "proper" headings and use 21 POS tags (with verb (VB) and gerund (VBG) being used only once). The most used tags are the same as in DMOZ, namely, they are NNP (25.35%), NN (22.89%), NNS (11.10%), JJ (8.81%), and CC, IN, CD (about 1% each). Unlike the DMOZ category names, the headings are more structured with commas (14.50%) and brackets (11.82%). We see that both datasets have similar part of speech tags distribution.

A qualitative analysis shows that the DMOZ category names are noun phrases, clearly divided into the "proper" and "common" categories. Looking at POS combination patterns, we can find 232 patterns describing the "common" categories. However, first twenty most used patterns (8.6%) describe already 99% of the "common" category names. Most of them are simple and structurally non-ambiguous, such as [NN] "Compensation" or [NN CC NN] "Pregnancy and Birth".

A qualitative analysis of LCSH headings reveals that they are chunks of noun phrases. The majority of them are in reverse order, such as "Dramatists, Belgian" [NNS, JJ]. There are also few naturally ordered examples, such as "Strikes and lockouts, Clothing trade" with the pattern [NNS CC NNS, NN NN], which can be simplified to two chunks [NP, NP].

A much wider set of POS tag patterns (13520) is needed to describe the subject headings. To cover 90% of the headings we need 1007 patterns (7.4%). To cover 95% of headings we need almost 3000 patterns (22.1%). However, if we look at the patterns at a chunk level (using commas as separators) we see 44 groups of chunk-patterns, where many chunks bear clear semantics. For example, the pattern [NNP NNP, NN CC NN, CD] seen at a chunk level transforms into [geo, NP, time], where "geo" stands for a geographical proper name, "NP" stands for a noun phrase, "time" stands for a time period. The example of a heading corresponding to this pattern is "United States, Politics and government, 1869-1877".

Parsing Labels with Simple Grammars

The parsing of labels in higher level structures can provide a better understanding of their semantic and thus to process them in a more meaningful notation for the computer. In particular, we want to use the SMatch algorithm [4] to align different classifications. This will allow, for example, the automatic translation of existing heterogeneous library classifications to a standard one.

The SMatch algorithm works on hierarchies of categories represented as lightweight ontologies [6] encoded in propositional description logic while the usual library classifications are represented in natural language. However, the use of the patterns discussed in the previous section can help us in translating the natural language labels in a formal language and improve the accuracy of our translation pipeline [5].

As a first step towards this process, we have developed a set of lightweight grammars for both datasets discussed in this paper. The grammar we constructed for DMOZ category

names contains ten production rules, expressed in Backus-Naur form and is recursive. This simple grammar already covers 99.81% of the category names. In comparison, the grammar for the LCSH subject headings contains seventeen production rules, is recursive and already covers 95.29% of headings.

These results show that a simple grammar can be used to parse accurately most of the patterns found in the state of the art classifications, thus providing extra understanding of the natural language without a loss in performance in the rest of the processing pipeline.

Conclusion and Future Work

In this paper we have first shown that a standard part of speech (POS) tagger could be accurately trained on the specific language of the metadata. A large scale analysis of the use of POS tags showed that the metadata language is structured in a very limited set of patterns that can be used to develop accurate lightweight Backus-Naur form grammars. We intend to use parsers based on these grammars to allow deeper understanding of metadata semantics, important for such tasks as translating classifications into lightweight ontologies for use in semantic matching.

In the future, we plan to simplify the grammars and try to unify them. We plan to analyse other metadata datasets and check the scalability of our approach and to investigate the possibility to automate the creation of grammar production rules.

References

1. P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. Journal on Data Semantics, 2005.
2. A. Doan and A. Halevy. Semantic integration research in the database community: A brief survey. AI Magazine, Special Issue on Semantic Integration, 2005.
3. B. Santorini. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania, 1990.
4. F. Giunchiglia, M. Yatskevich, P. Shvaiko. Semantic Matching: Algorithms and Implementation, Lecture Notes in Computer Science, Springer, 2007.
5. I. Zaihrayeu, L. Su, F. Giunchiglia, W. Pan, Q. Ju, M. Chi, X. Huang “From Web Directories to Ontologies: Natural Language Processing Challenges”, 6th International Semantic Web Conference, 2007.
6. I. Zaihrayeu, M. Marchese, F. Giunchiglia, "Encoding Classifications into Lightweight Ontologies". In: Proceedings of the 3rd European Semantic Web Conference 2006.
7. F. Giunchiglia, P. Shvaiko, M. Yatskevich, "Semantic schema matching". In: On the move to meaningful internet systems 2005: coopIS, DOA.
8. F. Giunchiglia, M. Yatskevich, P. Avesani, P. Shvaiko: “A Large Scale Dataset for the Evaluation of Ontology Matching Systems”, The Knowledge Engineering Review Journal (KER), Cambridge University Press, 2008.
9. F. Giunchiglia, P. Shvaiko, M. Yatskevich, "Discovering Missing Background Knowledge in Ontology Matching". In proceedings of 17th European Conference on Artificial Intelligance – ECAI, 2006.
10. J. Blitzer, R. McDonald, F. Pereira. "Domain adaptation with structural correspondence learning". Proceedings of the Empirical Methods in Natural Language Processing, 2006