On the Semantic Representation and Extraction of Complex Category Descriptors

André Freitas[1], Rafael Vieira[2], Edward Curry[1], Danilo Carvalho[3], João C. Pereira da Silva[2]

[1]Insight Centre for Data Analytics, National University of Ireland, Galway
[2]Computer Science Department, Federal University of Rio de Janeiro (UFRJ)
[3]PESC/COPPE, Federal University of Rio de Janeiro (UFRJ)

Abstract. Natural language descriptors used for categorizations are present from folksonomies to ontologies. While some descriptors are composed of simple expressions, other descriptors have complex compositional patterns (e.g. 'French Senators Of The Second Empire', 'Churches Destroyed In The Great Fire Of London And Not Rebuilt'). As conceptual models get more complex and decentralized, more content is transferred to unstructured natural language descriptors, increasing the terminological variation, reducing the conceptual integration and the structure level of the model. This work describes a representation for complex natural language category descriptors (NLCDs). In the representation, complex categories are decomposed into a graph of primitive concepts, supporting their interlinking and semantic interpretation. A category extractor is built and the quality of its extraction under the proposed representation model is evaluated.

1 Introduction

Ontologies, vocabularies, taxonomies and folksonomies provide structured descriptors for categories of objects and their relationships. While ontologies target a more centralised, consistent and structured representation of a domain, folksonomies allow a decentralised, less structured categorization. Both representation models have in common natural language descriptions associated with object categories. These *natural language category descriptors* (NLCDs) are a fundamental part of the communication of the meaning behind these artefacts. While some descriptors are composed of single words or simple expressions (e.g. 'Person', 'Country', 'Film'), other descriptors have more complex compositional patterns (e.g. 'French Senators Of The Second Empire', 'United Kingdom Parliamentary Constuituencies Represented By A Sitting Prime Minister').

As models get more complex and decentralized, more content is transferred to unstructured natural language descriptors, increasing the terminological variation, reducing the conceptual integration and the structure level of the model. In this scenario, the more formal conceptual model tools are substituted by complex NLCDs as an interface for domain description. From the perspective of information extraction and representation, NLCDs provide a much more tractable subset of natural language which can be used as an '*interface*' for the creation of structured domains. From the syntactic perspective, natural language category descriptors (NLCDs) are short and syntactically well-formed phrases. Differently from full sentences, NLCDs present simpler and more regular compositional patterns. By structuring NLCDs, we intend to support the creation of more structured resources with lower construction effort and in a more decentralized way.

In this work we describe a representation and an extraction approach for complex NLCDs. In the representation, complex predicates are decomposed into a graph of primitive word senses supporting the alignment between different NLCDs. A NLCD extractor is built and the extraction quality is evaluated. An extended version of this paper can be found at the website[1]

2 Related Work

Different works have focused on information and data extraction approaches applied in the context of semantic annotations and the Semantic Web. Most of these approaches have targeted the extraction of ontologies and datasets from semi-structured data [1], from unstructured data [7] or the alignment of folksonomies to ontologies [3][4][6]. YAGO [1] is a large-scale ontology which is automatically built from Wikipedia and WordNet. YAGO extracts facts from the infoboxes and the category system of Wikipedia, representing them in a data model which is based on reified RDF triples. YAGO builds a taxonomic structure from Wikipedia categories, aligning them to WordNet synsets. Specia & Motta [3] propose an approach for making explicit the semantics behind the tags, by using a combination of shallow pre-processing strategies and statistical techniques, together with knowledge provided by ontologies available on the Semantic Web. Cattuto et al. [4] proposed a systematic evaluation of similarity measures for folksonomies. Voss [6] concentrates on the description of an approach for the translation of folksonomies to Linked Data and SKOS. Comparatively, most of the previous works concentrate on the analysis and alignment of simple (non-complex) tags. Another difference is the proposal of a representation model which goes beyond a taxonomic structure.

3 Representation Model

The representation model is aimed towards facilitating the fine-grained integration between different NLCDs, providing the creation of an integrated and more structured model from the category descriptors. The representation also has an associated interpretation model which aims at making explicit the algorithmic interpretation of the descriptor in the integrated graph. A NLCD can be segmented into 7 representation elements:

Entity: Entities inside a NLCD are terms which are sub-expressions of the original category which can describe predications or individuals. The entities map to

[1] http://graphia.dcc.ufrj.br/nlcd

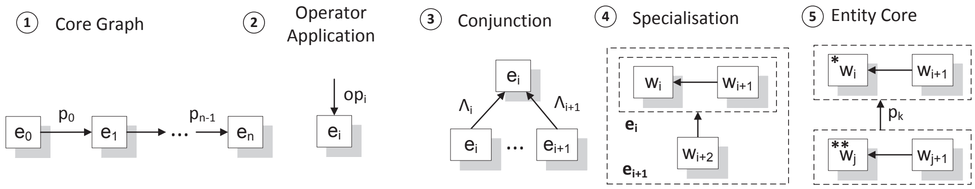

Fig. 1: Graph patterns showing the relations present in the graph representation.

a subset of the content words (open class words), which carry the main content or the meaning of a NLCD. Words describing entities can combine *nouns*, *adjectives* and *adverbs*. The entities for an example NLCD *‘Snow Or Ice Weather Phenomena’* are *‘Snow’*, *‘Ice’*, *‘Weather Phenomena’*. Entities are depicted as e_i in Figure 1(1).
Class & Entity core: Every entity contains a semantic nucleus, which corresponds to the phrasal head and which provides its core meaning. For the predicate *‘Snow Or Ice Weather Phenomena’*, *‘Phenomena’* is the class & entity core. Depicted as ‘*’ in Figure 1(5).
Relations: Relation terms are binary predicates which connect two entities. In the context of predicate descriptors, relation terms map to closed class words and binary predicates, i.e. prepositions, verbs, comparative expressions (*same as*, *is equal*, *like*, *similar to*, *more than*, *less than*). Depicted as p_i in Figure 1(1).
Specialization relations: Specialization relations are defined by the relations between words w_i in the same entity, where w_{i+1} is specialised by w_i. Represented by an unlabelled arrow in Figure 1(4).
Operators: Represents an element which provides an additional qualification over entities as a unary predicate. Operators are specified by an enumerated set of terms which maps to adverbs, numbers, superlatives, etc. Depicted in Figure 1(2).
Conjunctions & Disjunctions: A disjunction between two elements ($w_i \vee w_{i+1}$) over an element e_j is defined by the distribution of specialization relations: e_j is specialised by w_i and e_j is specialized by w_{i+1}. A conjunction is treated as an entity which names the conjunction of two entities through a conjunction labelled link. The conjunction representation is depicted in Figure 1(3).
Temporal Nodes: Consists in the representation of temporal elements references into a normalized temporal range format.

Further examples are depicted in Figure 2. The representation can be directly translated into an RDF (Resource Description Framework) graph. Most of the overhead in the translation is due to the fact that words mapping to classes need to be instantiated and later reified. Terms which are classes and which need to be reified are reflected as instances.

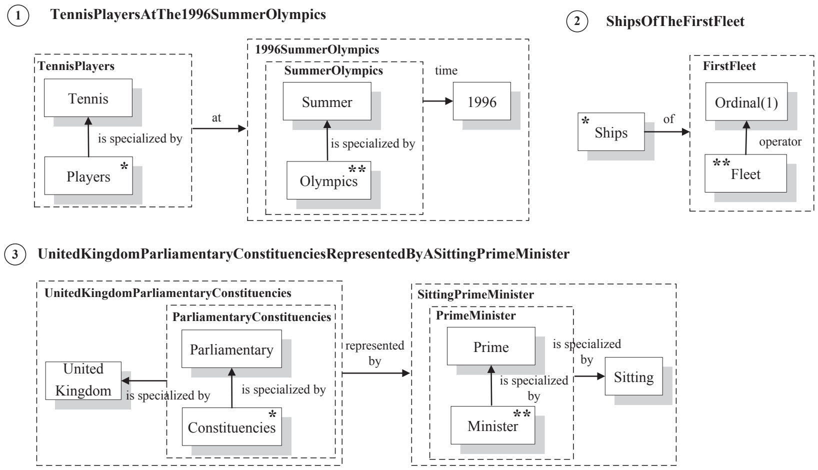

Fig. 2: Depiction of examples of NLCDs.

4 Extraction

This section describes the process for extracting NLCDs into the proposed representation model. Figure 3 shows the components and the extraction workflow. The NLCD extraction consists of the following steps:

POS Tagging: Detection of the lexical categories of the NLCD words. The extractor uses the NLTK POS Tagger[2].
Segmentation: The segmentation of the NLCD starts by detecting the relations and splitting the descriptor into a set of entities and relations.
Entity Detection: This step consists on the detection of 3 types of entities: named entities, operators and temporal references: (i) The detection of named entities is based on the creation of a gazetteer from DBpedia 3.9 instances. Elements tagged as nouns and proper nouns are checked against the gazetteer; (ii) Operators are detected using the combination of an enumerated list of operators and regular expressions based on POS Tags; (iii) Temporal references are detected using regular expressions and are normalized.
Specialization ordering: This step consists in defining the specialization sequence for the terms inside each entity. Two heuristic indicators are used in the determination of the ordering of the terms inside the classes: POS Tags and a corpus-based specificity measure (inverse document frequency (IDF) over Wikipedia 2013 text collection). The POS Tags are used to order the words based on the lexical categories. The ordering is defined by the relations (NN - *is specialised by* → JJ, JJ - *is specialised by* → RB). For an entity containing words from the same lexical category, IDF is used to define the ordering: Lower IDF - *is specialized by* → Higher IDF.

[2] http://www.nltk.org

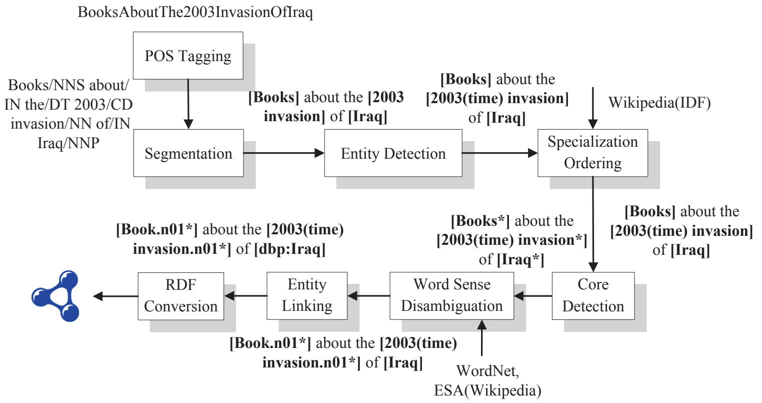
BooksAboutThe2003InvasionOfIraq
POS Tagging
Books/NNS about/ IN the/DT 2003/CD invasion/NN of/IN Iraq/NNP
Segmentation
[Books] about the [2003 invasion] of [Iraq]
Entity Detection
[Books] about the [2003(time) invasion] of [Iraq]
Wikipedia(IDF)
Specialization Ordering
[Books] about the [2003(time) invasion] of [Iraq]
Core Detection
[Books*] about the [2003(time) invasion*] of [Iraq*]
Word Sense Disambiguation
WordNet, ESA(Wikipedia)
[Book.n01*] about the [2003(time) invasion.n01*] of [Iraq]
Entity Linking
[Book.n01*] about the [2003(time) invasion.n01*] of [dbp:Iraq]
RDF Conversion

Fig. 3: Extraction components and workflow.

Word Sense Disambiguation (WSD): The WSD component is used to align the extracted words with their WordNet senses, based on the context in which the word occurs (the NLCD). Let the sequence of words $w_0, w_1, ..., w_n$ be the natural language descriptor for a category c. Let $g_0, g_1, ..., g_k$ be the WordNet glosses associated with the senses for w_i for $0 \leq i \leq n$. Let $\kappa(w_i)$ be the context of w_i defined by $w_0, w_1, ..., w_n$ (excluding the target word). The sense for w_i is given by $sr_{ESA}(\kappa(w_i), g_j)$ where sr_{ESA} is the distributional semantic relatedness measure (Explicit Semantic Analysis) between the WordNet glosses and the category context.
Entity linking: The Entity linking component aligns terms in the extracted graph to DBpedia entities. Entity linking uses DBpedia as a named entity base and a ranking function based on TF/IDF over labels, entity cardinality and levehnstein distance.
RDF conversion: At this point the relations are represented as an internal set of extracted graphs following the proposed representation model. The model is then converted into an RDF graph.

5 Evaluation

The extraction approach was evaluated by randomly selecting a sample of 2,696 Wikipedia categories from the original set of 287,957 categories. These categories were extracted and manually evaluated according to eight extraction features (Table 1). The features map to the components of the extraction approach. Table 1 shows the accuracy for each feature.

The low error in entity segmentation, relation extraction and specialization sequence shows the generality of the extraction rules in relation to the tractable subset of natural language category descriptors. Additionally, the high accuracy in the determination of the sequence of specialization relations, detection of class

	Entity Segmentation	**Relations**	**Unary Operators**	**Specialization Relations**	**Class Core**	**Entity Core**	**WSD**	**Entity Linking**
Accuracy	79.38%	95.96%	99.74%	97.81%	99,37%	81.86%	82.2%	78.1%

Table 1: Accuracy for each extraction feature.

and entity cores shows the correctness in the construction of the taxonomic structure. For an open domain scenario, the WSD approach based on Explicit Semantic Analysis achieved an average accuracy of 82,2%.

Additionally, the graph extraction time was evaluated with regard to the extraction performance time. The experiment was carried in a 1.70GHz CPU computer with 4GB RAM. The extraction was evaluated with regard to three main categories: (i) graph extraction time (**9.8 ms per graph**), (ii) word sense disambiguation **121.0 ms per word** and (iii) entity linking **530.0 ms per link**. The overall extraction time per NLCD shows that the approach can be integrated into medium-large scale categorization tasks. Each category generates an average of 10.2 RDF triples. The extraction tool is available at the website[3].

6 Conclusion & Future Work

This paper analyses the use of complex natural language category descriptors (NLCDs) and proposes a representation model and an extraction approach for NLCDs. The accuracy of the proposed approach was evaluated over Wikipedia category links, achieving an overall structural accuracy above 78%. Future work will focus on the evaluation of the approach under domain-specific NLCDs.

References

1. Suchanek, F. Kasneci, G., Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet, In Proc. of the 16th Intl. Conf. on World Wide Web, pp. 697-706 (2007).
2. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In Proc. of the Intl. Joint Conf. On Artificial Intelligence (2007).
3. Specia, L., Motta, E.: Integrating Folksonomies with the Semantic Web: In Proc. of the 4th European Conf. on the Semantic Web, pp. 624-639, (2007).
4. Cattuto C., Benz, D., Hotho, A., Stumme, G.: Semantic grounding of tag relatedness in social bookmarking systems. In Proc. 7th Intl. Semantic Web Conference (2008).
5. Limpens, F., Gandon, F., Buffa, M.: Linking Folksonomies and Ontologies for Supporting Knowledge Sharing: a State of the Art, Technical Report (2009).
6. Voss, J.: Linking Folksonomies to Knowledge Organization Systems, Communications in Computer and Information Science v. 343, pp 89-97 (2012).
7. Cimiano, P., Handschuh, S., Staab, S.: Towards the Self-Annotating Web, In Proc. of the 13th Intl. Conf. on World Wide Web, pp. 462-471 (2004).

[3] http://graphia.dcc.ufrj.br/nlcd